THE UNIVERSITY OF MELBOURNE

Data, Systems and Society Research Network (DSSRN)

Data and inequity:
Who's missing in big data?

Edited by Ruth De Souza

Edited by Ruth De Souza

Copyediting by Abbra Kotlarczyk

Images by Debris Facility PTY Ltd

Copyright rests with the contributors.

The contents of this publication excluding the introduction have been peer-reviewed.
We thank the following people: Ximena Comacho, Donna Cormack, Angela Daly, Julie McLeod, Gerard Goggin, Monique Mann, Jack Nunn, Michael Rigby, Liz Sonenberg, Chen Zong

ISBN 978-0-7340-5596-5

This research was funded by DSSRN

TABLE OF CONTENTS

ABOUT DSSRN

Ruth De Souza, Rob Moss

The Data, Systems and Society Research Network (DSSRN) was established in 2015 in response to the growing role of data in research across the University of Melbourne. DSSRN reflects a growing trend for universities to support cross-disciplinary collaborations in the belief that combining methods and data from different fields can generate novel solutions to complex problems. This network shares knowledge, tools and resources in the broad area of data, systems, and society across the University.

HISTORY

DSSRN was formed in 2015, as a consortium arising from three applications to the University of Melbourne's Hallmark Research Initiatives program, which supports interdisciplinary research communities across the University by harnessing cross-University capabilities and increasing the impact of high quality research.

Each initiative addressed a complementary aspect of Data Science:

- Hallmark initiative in Data Science – proposed to link disparate research disciplines and commercial enterprises with expertise in Data Science research infrastructure and methods, to cross-pollinate across tools and techniques;
- Hallmark initiative in Complex Social Dynamics – proposed to link researchers and technology developers to accelerate theory informed knowledge on complex systems;
- Hallmark initiative in Merging Perspectives in Human Population Sciences – proposed to merge perspectives, information and skills across qualitative and quantitative human sciences to avoid errors associated with inferring one from the other.

ENHANCING DATA SCIENCE RESEARCH CAPABILITIES.

DSSRN launch

The network was launched on Wednesday 14 September, 2016 and brought together over a hundred students, academics and support staff from across the University and people working in the fields of big data, data science and informatics from the Victorian Government. It provided an opportunity for attendees to:

- Form new interdisciplinary collaborations;
- Discover available data sets and tools; and
- Define a research agenda for building University capacities in these areas.

The launch included 'Sound bite' talks, 'Sprint Sessions' about modelling research data and telling stories with data. Participants were invited to collaboratively assemble a 'wish list' of data techniques, capabilities, and infrastructure. These were supported by a range of data visualisations, research platforms, and other resources displayed to showcase the diverse range of capabilities and activities across the University at lunch. The day concluded with a session on Capability Mapping facilitated by Mark Fallu which explored questions including: What are our current capabilities? Where are the gaps? How should we go about addressing them? How can we best identify new collaborations? Topics and themes for future DSSRN events were collated informed by feedback collected throughout the day, and key aspects of a research agenda for building University capacities were compiled.

DSSRN reflects a growing trend for universities to support cross-disciplinary collaborations in the belief that combining methods and data from different fields can generate novel solutions to complex problems

Figure 1: Early DSSRN wishlist of data techniques, capabilities, and infrastructure 2016

Garage sale

In November 2016, DSSRN had its first data garage sale which involved sharing knowledge, resources and capabilities. Researchers and professional staff attended from across the University. Knowledge was shared about the complexity of data sets, available methods for analysing data and resources available for researchers to develop their skills in working with data. This event highlighted that the people most likely to attend events such as these, are those who: (a) already have these kinds of collaborations in place; and/or (b) work partly or primarily on methods development. They are positively disposed to this kind of activity. In contrast, people who collect and hold data were more likely to be risk-averse and to perceive data-sharing as including some element or risk of "loss". For some of these people, this is (unfortunately) based on prior negative experiences of collaborating with people outside of their own discipline.

Consultation

A University-wide consultation undertaken by DSSRN identified several priorities for advancing data-driven research at the University of Melbourne including: Facilitating access to data infrastructure and platforms; developing research capacity and training; promoting research collaborations, and providing a focal point for external engagement.

Equity and data symposium

In November 2018, DSSRN hosted a successful symposium on data and equity, networking approximately 200 academics and providing a University-wide platform for considering the interdisciplinary impacts of data in research and practice. This was an effort to engage with research communities with emerging data science needs as well as to engage data scientists in ethical questions. It was envisaged that this theme would bring together potential collaborators across discipline domains to develop potential research agendas and identify opportunities afforded by partnerships, emerging technologies and approaches.

Knowledge sharing

People in different disciplines or research areas often have much in common, but are unaware of this commonality due to many factors including (a) limited exposure to people outside of their own disciplinary area of expertise; (b) field-specific terminology and practices that cloud this commonality; (c) time pressures that may prevent/restrict engagement in exploratory activities (e.g., establishing new interdisciplinary collaborations) in the absence of guaranteed success.

DSSRN fosters collaboration and engagement by bringing together domain experts and methodological experts from all parts of the University who otherwise rarely cross paths. Enabling the identification of shared opportunities and building understanding of the University's capability in the Data domain.

DSSRN members and participants at DSSRN events repeatedly remarked that the most successful or rewarding collaborations often arose from unexpected encounters and conversations, rather than from structured projects or plans. Many researchers learned new things about people, capabilities, and activities within their own department via DSSRN meetings and events — the limited awareness of common interests, challenges, and solutions is not restricted to people in different schools or faculties.

Staffing
Professor Jodie McVernon has led DSSRN as the Chair since its inception. The network has been staffed by two part-time academics from diverse discipline domains to act as knowledge brokers in order to define priority areas of focus for formal community building activities and networking events. In the first two years Dr Rob Moss and Mr Suneel Jethani were the Academic Convenors and in 2018 Dr Gideon Aschwanden and Dr Ruth De Souza became co-convenors. Staff at the Networked Society Institute (NSI) especially Adam Lodders, Kate Murray and Fiorella Chiodo have provided support and advice to the DSSRN team throughout.

Steering committee
DSSRN has had an active and engaged Steering Group, who have assisted with the dissemination of knowledge and promotion of events as shown by the breadth and (academic) diversity of the committee. Members have included representatives from all faculties: Architecture, Building and Planning; Arts; Business and Economics; Education; Engineering; Fine Arts and Music; Law; Medicine, Dentistry and Health Sciences and Science.

The future
DSSRN completes its tenure as a virtual network this year, it has provided an opportunity for conversations on data and research that were valuable and different from other entities in the University, by combining traditional and nontraditional data scientists. DSSRN has helped to shape and inform the University's major investment into the Petascale Campus Initiative (PCI) where world class hardware and investment in the on ramp of world class people provides an opportunity for people to acquire hardware and people skills. However, by definition the PCI focus has been on internal infrastructure, and there are still many challenges and hurdles to effectively supporting interdisciplinary collaborations at the University that reach out to the broader community, and provide a shared understanding of the university's capability in the data domain.

DSSRN fosters collaboration and engagement by bringing together domain experts and methodological experts from all parts of the University who otherwise rarely cross paths

DATA AND EQUITY: THE SYMPOSIUM

by Ruth De Souza

DATA AND EQUITY: THE SYMPOSIUM

Contemporary life is datafied; data are aggregated from sources as diverse as health records, advertising, retail activities, social media timelines and public records. Sophisticated computational techniques such as big data analytics and machine learning have become central to ordering our lives. The ubiquity of big data research methods, greater computing power and larger and more complex datasets, have created the potential for speed, efficiency and novel insights into analysis of social issues. However, the seemingly value-neutral capacity to correlate large data sets to produce valuable knowledge and to automate decision making also present urgent ethical questions (Daly et al., 2019; Elmer et al., 2015 and Zook et al., 2017). In this complex ecosystem of pervasive data collection—what we might in shorthand call *data surveillance*—data-intensive social transformation can further entrench harm, particularly in societies already marked by inequity (Eubanks, 2018). Furthermore, the combination of technical complexity and corporate secrecy means that the opacity of both algorithms and the data shaping them are 'black-boxed,' leaving the public with little recourse for potential harms (Pasquale, 2015). Consequently, how these systems transform social inequalities and power differentials in a data-enabled society is of important scholarly concern.

Critical assessment of the effects of big data practices are emerging in a number of fields and cohorts, including the use of predictive data-driven risk assessment tools in welfare and policing (O'Neil, 2016); racial bias in Google's algorithms (Noble, 2018); and the surveillance and management of racialised populations (Metcalfe & Dencik, 2019; and Taylor, 2016). Further, strategies for enhancing individual and collective control of data—and protecting marginalised groups from what has been termed 'informational imperialism' and 'digital colonialism'—are beginning to emerge (Broad, 2018 and Mann & Daly, 2018). For example, Indigenous scholars and communities are developing concepts of Indigenous Data Sovereignty (IDS) and Indigenous Data Governance (IDG) to prevent and ameliorate data colonialism (Kukutai & Walter, 2017; Lovett et al., 2019; Walker, Lovett, Kukutai, Jones, & Henry, 2017). A recent edited book on 'Good Data' (Daly & Mann, 2019) attempts to broker a multi-disciplinary and multi-stakeholder conversation on how digital technologies and data can enable human flourishing.

A gap in many data initiatives has been their focus on resource-intensive research computing infrastructures and computational, data and informatic capabilities at the expense of considering ethical issues including the use of personal information; access and ownership of data sets; and the impacts of research outcomes (Shorey & Howard, 2016). Critiques of datafication and "digital positivity" and their impacts on individuals and communities are emerging which also consider frameworks for ensuring these rapid developments are ethical and anti-oppressive. Strategies include integrating feminist and postcolonial science studies with an ethics of care; data justice, human rights approaches and non discrimination (Leurs, 2017; Milan & Treré, 2019). More recently Cifor, et al (2019) have developed the Feminist Data Manifest-No which articulate a set of commitments for feminist data studies. At the Data Justice Lab at Cardiff University's School of Journalism, Media and Culture, scholars are interrogating the relationship between datafication and social justice, specifically focusing on the politics and impacts of data-driven processes and big data. (Dencik et al., 2019; Heeks & Shekhar, 2019; and Metcalfe & Dencik, 2019). The independent, nonprofit research institute Data & Society in New York aims to advance public understanding of the social implications of data-centric technologies and automation. Its focus is on exploring whether human rights-based frameworks are applicable to individuals developing and deploying artificial intelligence (AI) and how fairness and accountability can be encoded into technical systems. In 2018 Human Rights Watch and a coalition of rights and technology groups developed the Toronto Declaration on human rights standards for machine learning, to ensure machine learning applications respect the principles of equality and non-discrimination.

With this context in mind, The Data, Systems and Society Research Network (DSSRN)—a collaborative research network at the University of Melbourne—organised a symposium bringing experts from a range of disciplines and domains together to identify key issues in equity and datafication through the four themes of i) Indigenous people; ii) cities; iii) health; and iv) ethics and privacy. As organisers of the symposium, we were tasked with encouraging speakers in each of these four panels to consider how digital data could amplify or create new kinds of inequities, and what kinds of mechanisms could be put in place to ameliorate inequity.

The symposium structure

The symposium began with the Indigenous cultural protocol of a Welcome to Country by Wurundjeri Elder Aunty Diane Kerr who identifies with the Ganun Willam Balak clan.

The official welcome was followed by an introduction to the day by Dr Ruth De Souza, with networking activities using visualisation techniques being facilitated by Dr Gideon Aschwanden and Ishita Chatterjee. This activity was established to encourage researchers working in related areas to gather throughout the day.

Howard Bondell, Professor of Statistical Data Science and Fellow of the American Statistical Association, began with a striking example of how histories of discrimination can become part of the statistical logic of everyday algorithmic systems, particularly if they are reproduced without human reasoning and/or intervention. He cited the defeat of presidential candidate and Republican nominee Gov. Thomas Dewey on November 2, 1948, an outcome which contradicted the polls predicting a Dewey victory over President Harry S. Truman. As is now well known, the people who had carried out the polling

on this particular occasion had chosen neighborhoods they were personally comfortable with. Consequently, the statistical findings were not generalisable. Professor Bondell shared examples from personalised medicine, hot-spot policing and the automation of job applications to advocate for how the design and training of machine-learning systems must be carefully considered to avoid bias being built into AI. Bondell concluded by challenging researchers to think hard about: What data are being used and for what purpose; how and from whom data are being collected; the apparent limitations of data; and how corrective measures might be made.

The first panel of the symposium, *Data and Indigenous People*, was chaired by Dr Tess Ryan and featured panellists Professor Marcia Langton (AM), Dr Lyndon Ormond-Parker and Darren Clinch. Langton outlined the aspirations of the Indigenous Data Network based at the Indigenous Studies Unit, University of Melbourne. In context of the Australian Government's failed targets for the Closing the Gap campaign, implemented to improve Indigenous health, education and social participation, Langton emphasised the urgent need for innovative and paradigmatic shifts in how to address Indigenous disadvantage. A focus on Indigenous data sovereignty can challenge national 'deficit thinking' approaches; it can allow Indigenous people to set their own agendas by using their own data—that is, data about themselves—to help secure the social, economic, cultural and health-based futures of their communities. Dr Lyndon Ormond-Parker's presentation examined the tensions in repatriating digital data in remote Aboriginal community archives (Wadeye) and the preservation of community languages, history and culture. In particular, the paradox of relying on recordings taken by settlers to learn about Aboriginal culture. Darren Clinch, our final speaker spoke about innovative data practices which combine Aboriginal art and technology practices and promote Indigenous Data Sovereignty and Governance principles and practices.

Dr Gideon Aschwanden facilitated the second panel, *Data and Cities*, which featured panellists Professor Michele Acuto, Dr Soheil Sabri and Ishita Chatterjee. Collectively they considered how the increasing number of sensors and storage has provided an unprecedented volume of information about cities, ranging from satellite images to real-time transportation flows. Such data has optimised numerous systems and improved the way cities are managed, planned, built and envisioned. Identifying issues of inconsistent depth and/or quality, the panel discussed the advantages and problems of data driven urbanism in local and global contexts, specifically the problem of 'missing people' in data and how to address it.

The *Data and Health* panel chaired by Professor Jodie McVernon with panellists Professor Anne Kavanagh, Professor Karin Verspoor and Associate Professor Steven Tong followed. Kavanagh explored the democratisation of disability data in the context of the National Disability Insurance Scheme (NDIS). Costing the Australian Government approximately $22 billion annually and providing critical support to people with disability, their families and carers, the NDIS allows limited access only to its key outcome data. Verspoor provided providing commentary on the value of sharing data and Australia's controversial electronic 'My Health Record,'. Verspoor examined the various favourable and missed opportunities and potential pitfalls of health record data—particularly in the context of secondary use of data—to enquire about academic researchers' responsibilities in respect to protecting individual privacy. Tong concluded with an in-depth exploration of the burden of skin infections and using old and new data for improving skin health.

Ethics, privacy and security are the most extensively discussed concerns in relation to big data uses. The final panel of the day, *Ethics and Privacy*, was chaired by Dr Fiona Tweedie with panellists Professor Simon Dennis, Dr Vanessa Teague and Professor Julie McLeod. This panel considered social, political and technological uses and effects of large-scale data in the twenty-first century. In prompting the question of 'who is missing?' in Big Data. The panel asked questions such as: What are the ethical debates that matter? and what still needs to be considered beyond the basic codes and protocols for the governance of data?. The relevance of concepts such as justice, empathy, agency, ownership, subjectivity and identification to data were central to this final panel.

The day concluded with a discussion between Dr Ruth De Souza, Darren Clinch, Dr Gideon Aschwanden and Professor Jodie McVernon.

The increasing number of sensors and storage has provided an unprecedented volume of information about cities, ranging from satellite images to real-time transportation flows. Such data has optimised numerous systems and improved the way cities are managed, planned, built and envisioned.

REFERENCES

Broad, E. (2018). *Made by humans: The AI Condition*. Melbourne: Melbourne University Press

Cifor, M., Garcia, P., Cowan, T.L., Rault, J., Sutherland, T., Chan, A., Rode, J., Hoffmann, A.L., Salehi, N., Nakamura, L. (2019). Feminist Data Manifest-No. Retrieved December 16, 2019, from Feminist Data Manifest-No website: https://www.manifestno.com/

Daly, A., Devitt, K., & Mann, M. (Eds.). (2019). *Good data*. Amsterdam: Institute for Network Cultures.

Dencik, L., Hintz, A., Redden, J., Treré E. (2019). Exploring data justice: Conceptions, applications and directions. *Information, Communication & Society.* 22(7):873–881. doi:10.1080/1369118X.2019.1606268

Elmer, G., Langlois, G. and Redden, J. (2015) *Compromised data: From social media to big data*. London: Bloomsbury

Eubanks, Virginia. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* New York: St. Martin's.

Heeks, R., & Shekhar, S. (2019). Datafication, development and marginalised urban communities: an applied data justice framework. *Information, Communication and Society*, 22(7), 992–1011. https://doi.org/10.1080/1369118X.2019.1599039

Kukutai, T., & Walter, M. (2017). Indigenous Statistics. In P. Liamputtong (Ed.), *Handbook of Research Methods in Health Social Sciences* (pp. 1–16). https://doi.org/10.1007/978-981-10-2779-6_40-1

Leurs, K. (2017). Feminist data studies: Using digital methods for ethical, reflexive and situated socio-cultural research. *Feminist Review*, 115(1), 130–154. https://doi.org/10.1057/s41305-017-0043-1

Lovett, R., Lee, V., Kukutai, T., Cormack, D., Rainie, S., & Walker, J. (2019). Good data practices for Indigenous data sovereignty and governance. In A. Daly, S. K. Devitt, & M. Mann (Eds.), *Good data* (pp. 26–36). Amsterdam: Institute of Network Cultures.

Mann, M., & Daly, A. (2018). (Big) Data and the North-in-South: Australia's informational imperialism and digital colonialism. *Television & New Media 20* (4): 379–95. https://doi.org/10.1177/1527476418806091

Metcalfe, P., & Dencik, L. (2019). The politics of big borders: Data (in)justice and the governance of refugees. *First Monday*, 24(4). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/9934/7749

Milan, S., & Treré, E. (2019). Big data from the South(s): Beyond data universalism. *Television & New Media, 20*(4), 319–335. https://doi.org/10.1177/1527476419837739

O'Neil C (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy.* New York: Crown Publishing Group.

Pasquale, F. (2015) The black box society: *The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Shorey, S., & Howard, P. N. (2016). Automation, big data and politics: A research review. *International Journal of Communication*, 10(0), 5032–5055.

Taylor, L. (2016). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning. D, Society & Space*, 34(2), 319–336. https://doi.org/10.1177/0263775815608851

Walker, J., Lovett, R., Kukutai, T., Jones, C., & Henry, D. (2017). Indigenous health data and the path to healing. *The Lancet*. Retrieved from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)32755-1/abstract?code=lancet-site

Wells, M. T., Ajunwa, I., Barocas, S., Duffy, B. E., & Ziewitz, M. (n.d.). *Algorithms, Big Data, and Inequality: ISS Collaborative Project Proposal 2018-2021*. Retrieved from http://socialsciences.cornell.edu/wp-content/uploads/2018/03/ISS-Algorithms-Public.pdf

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J., Narayanan, A., Nelson, A. and Pasquale, F., 2017. Ten simple rules for responsible big data research. *PLoS Computational Biology*, 13(3), pp. 1–10. doi:10.1057/s41305-017-0043-1

OVERVIEW OF SYMPOSIUM THEMES

by Ruth De Souza

This interdisciplinary collection of peer-reviewed papers from the Data, Systems and Society Research Network (DSSRN) Symposium aims to expand academic knowledge around questions of equity and data across three key domains: Indigenous people, cities and health. The fourth panel considers ethical frameworks for navigating issues of inclusion, exclusion and surveillance.

For this publication, graduate researchers at the University of Melbourne have synthesised key issues raised by the four panels of experts. These are prefaced a piece by Dr Fiona Tweedie titled *Ethics, Justice and the Problem of Data* which examines opportunities to consider how digital technologies and data can be used productively and justly for all communities.

Tweedie outlines the limitations of conducting data-intensive research, and interrogates what is at stake when data is reused and analysed beyond its original context. Tweedie notes the importance of ensuring that people who are the most likely to be affected—that is already-marginalised groups—are included in proposed remedies, especially given that research in artificial intelligence and machine learning happen in computational science departments and technology companies that do not typically foster these relationships. Tweedie proposes that analytic methods from the humanities, arts and social science (HASS) disciplines can be drawn upon as they already engage in data about humans. In asking for algorithmic accountability, Tweedie critiques the limitations of normative ethical frameworks for conducting data-intensive research, where little guidance is available about creating, using and sharing large datasets. Tweedie advocates for models of social justice to be used to hold both algorithms and their creators accountable for the outcomes of their work.

As the scale and complexity of data expand, complex ethical debates at the centre of big data provide unprecedented challenges for scholars. In *Beyond Bigness: Can Big Data Have An Ethical Future?* Tyne Daile Sumner considers the shifting definitions of privacy and data protection rights in an age of big data. Defining big data beyond the catchy notions of "volume, variety and velocity", Sumner includes "behaviours, practices, networks, infrastructures and politics" that challenge pre-existing moral and ethical norms. In this data-saturated context, Sumner challenges the illusion of big data as neutral and asks for vigilance on behalf of people and groups who are marginalised. Raising questions about representation, bias, marginalisation and hypersurveillance. Sumner concludes by considering the social, political and technological uses and effects of large-scale data by asking: Who is missing? What are the ethical debates that matter? and, What still needs to be considered beyond the basic codes and protocols for the governance of data?.

In *Data and Indigenous People*, Amba-Rose Atkinson outlines how data has been captured and used to entrench a discourse of deficit which permeates the experiences of Indigenous people in Australia. Atkinson considers issues including the reliability of data, data literacies, data protection, and the role of data in maintaining cultural knowledge. Indigenous data sovereignty signifies an attempt to regain ownership, control, and distribution of such data. Atkinson argues that without cultural and intellectual property protections, historically sensitive Indigenous knowledges were extracted to further non-Indigenous ends with little accountability for how this knowledge was acquired, used and profitted from. Many scholars view contemporary data practices as new forms of colonialism; against this backdrop, Atkinson concludes with a summary of the robust discussion facilitated by Dr Tess Ryan, which featured Professor Marcia Langton, Dr Lyndon Ormond-Parker and Darren Clinch.

Ishita Chatterjee and colleagues consider the proliferation of urban data, ranging from frequently updated satellite imagery to real-time transportation data. This data provides the basis for policy formulation, drafting development plans and modelling future scenarios. As they make apparent, however, this data is not available everywhere in the same depth or quality and is often fragmented and/or unreliable. Members of the Data and Cities panel—Dr Gideon Aschwanden, Dr Soheil Sabri, Professor Michele Acuto and Ishita Chatterjee—who are each experts in urban analytics, policy, science and informal settlements, and the limits and opportunities of data-driven urbanism review these issues.

Kyle Turner, Professor Jodie McVernon and Ruth De Souza report on three advances in data and health. The introduction of the National Disability Insurance Scheme (NDIS) signals a universal rights-based scheme representing a historic shift in allocation principles within Australia's disability policy sector. The scheme aims to provide lifelong, individualised support for people living with disability. Greater access to data for researchers is needed, however, in order to strengthen public health policy and ensure effectiveness and accountability. The second issue that Turner and McVernon address relates to the potential benefits and pitfalls of Australia's new electronic My Health Record—an electronic personal health record system that was introduced in July 2012. By August 2017, approximately 21 percent of Australia's population had registered to use My Health Record amidst concerns about access and privacy. Finally, Turner, McVernon and De Souza report on the remarkable potential of linking data, where dozens of smaller data sets are combined and analysed to reveal new insights and public health benefits that would have otherwise remained hidden.

Collectively, these contributions provide a range of viewpoints and settings from which to consider data and equity. Through local engagement with various issues of concern, these symposium reports highlight the challenges and opportunities of datafication through the lenses of Indigenous people, health, cities and ethics. These considerations move beyond the technical to consider, the social and cultural implications of datacentrism. Many of the researchers within this collection provide empirical analyses of the ways different groups of people can be disparately affected by data use. It is our hope that by collating these papers we might stimulate conversations and research collaborations both within and beyond the university, in order to consider how and where data can empirically and conceptually advance societal aspirations for flourishing.

ETHICS, JUSTICE AND THE PROBLEM OF DATA

by Fiona Tweedie

The increasing availability of large datasets and the development of the computational tools with which to analyse them has opened up new possibilities for research, both within the academy and in governmental and commercial entities. These new possibilities, however, bring with them challenges that are testing the frameworks of research practices and are demanding the development of new methodological approaches. When computational approaches are brought to bear on datasets with insufficient care afforded to the context in which the data was created, or a lack of attention to the consequences of errors, the results vary widely. Outcomes can range from the naïve—in the case of insufficiently rigorous research[1]—to the disastrous—when humans fall victim to automated decision-making without adequate review or appeal options.[2] The problems facing anyone conducting data-intensive research are two-fold. The first question concerns the development of methodologies that are sensitive to the challenges of working with large assemblages of data, particularly when they are divorced from their original context of collection. Due to the necessity of using computational tools to process significant volumes of data, much of the research in data science—in particular Artificial Intelligence (AI) and machine learning—occurs within computational science departments and technology companies. Here, methods from the Humanities, Arts and Social Science (HASS) disciplines can be drawn upon, as these fields, especially the Social Sciences, are already deeply engaged with how to approach data about humans. The second question is one that affects HASS disciplines as much as their counterparts in Science, Technology, Engineering and Medicine (STEM) and concerns the ethics of conducting data-intensive research. Statements of research ethics, such as the Australian *National Statement on Ethical Conduct in Human Research* (National Health and Medical Research Council (NHMRC), 2018b), recognise that data about humans must be handled responsibly, but are often silent on the specific difficulties posed by creating, using and sharing large datasets, such as those generated from social media posts and other online activities. I will argue in the following that models of social justice can be used to help address these shortcomings. Rather than seeing 'ethics' as a permission-seeking exercise, the linking of ethics to justice requires that practitioners—whether holding academic appointments or not—should engage with the questions of who is present in their data and how they might be affected in both social and economic terms. By bringing methods developed within fields such as anthropology to bear on data usage, and by examining these impacts in terms of justice, I believe that it is possible to identify means of conducting research that is both methodologically sound and socially responsible.

1. For instance, a network analysis of the Odyssey conducted by three physicists (Miranda, Baptista & Pinto, 2018) attracted criticism for failing to consult any significant scholarship on Homer, leading to unsupportable conclusions about the nature of Bronze Age Society (Gainsford, 2018). This exercise demonstrates that attempts by researchers to apply analytical methods across disciplines without due regard to context will runs the risk of falling into basic errors.
2. Broad (2018) provides an excellent overview of the Australian 'Robodebt' disaster, in which automated systems at Centrelink targeted individuals for welfare debts, often incorrectly, leading to huge distress and widespread criticism of both the policy and its implementation.

Rather than seeing 'ethics' as a permission-seeking exercise, the linking of ethics to justice requires that practitioners—whether holding academic appointments or not—should engage with the questions of who is present in their data and how they might be affected in both social and economic terms.

UNDERSTANDING DATA

The availability of large assemblages of data has raised questions of how such datasets should be approached and analysed. One of the temptations of big data[3] lies in its promise that, due to its size and complexity, it comes closer than small datasets to representing reality. Mauthner (2018, p. 21) characterises the positivist attitude that accompanies research using big data as "the widespread belief that large datasets, combined with computational techniques [...] reveal their inherent truths." Additional challenges arise from the practice of publishing datasets to enable other researchers to reuse and reanalyse them. There are two significant problems with this faith in data as a reusable resource. The first issue stems from the fact that any dataset, no matter how large, is produced and analysed by humans, invalidating the assumption that the larger the dataset the more inherently reliable it is. The second is a deeper question of what it means to assemble a dataset. In grappling with these questions of how to approach a dataset, thinking developed within the social sciences about research methodologies and the role of the researcher can bring nuance to the data positivism described by Mauthner above.

I will begin with the smaller of these questions—the presence of human practitioners in assembling and analysing datasets. In a study of data sharing practices, Mauthner and Parry (2013, p. 58) investigate the premise that they see as underlying modern research data sharing schemes, namely that datasets are "separate, rather than inseparable, from the contexts that generate them." The case can be made that research datasets—perhaps most readily those from the natural sciences—can be reused for novel purposes by researchers who are removed from the original purpose of the data collections. An example of successful reuse of a dataset is found in Fisher's Iris dataset. This dataset, first published in 1936, describes the morphology of 150 iris flowers, fifty from each of three species. This dataset has since gone on to be reused widely for purposes including training machine learning classifiers, a use that Fisher could scarcely have anticipated. The status of the Iris dataset as the basis of many subsequent and unrelated analyses would seem to make the case for making data available for reuse. Use of existing datasets to train machine learning classifiers does not, however, always proceed so smoothly. The Iris dataset was originally collected as part of Fisher's statistical work developing linear discriminant analysis. It is, then, particularly suitable for use in machine learning as the clustering of data into three species is well understood. A counter-example is found in Broad's (2018, pp. 10–11) account of Oakden-Rayner's reservations of the proposed use of the ChestX-ray14 dataset to train artificial intelligences to recognise signs of lung disease. Unlike the carefully curated Iris dataset, this dataset consists of 112,120 chest X-rays of 30,805 individuals collected from patients in the USA. The labels for the fourteen pathologies of interest have been extracted from the dataset via text mining (Wang et al., 2017). At the root of Oakden-Rayner's (2017) critique of using this dataset to train an artificial intelligence is the objection that the labels were not created with this purpose in mind. He found that there were inaccuracies in the labels used. Even worse for the training of an artificial intelligence was the fact that some features—such as the presence of a chest drain—had not been labelled, since they are sufficiently obvious to a human clinician to not require comment (Oakden-Rayner, 2017). This lack of a label creates the risk that the artificial intelligence will learn to look for the chest drain and not the pathology it is supposed to be identifying. While Oakden-Rayner (2017) stresses his hopes for the application of deep learning to medical imaging, the example of the chest X-ray dataset indicates that data must be fit for purpose. In this case, training AIs requires clean, consistently structured data to ensure that the results are meaningful.

Data collection is seldom undertaken without a purpose. Mauthner and Parry (2013, p. 58) argue that, across disciplines, data collection is "tied to the specific scientific questions, contexts and projects [that researchers] are working on." In the case of data collected as part of social science fieldwork, the relationship between data and researcher is more overtly interpersonal and contextual than in the natural sciences. Indeed, in the case of interview-based work, the relationship between the researcher and their subject can influence the information disclosed. For instance, McLeod and Thomson (2009, pp. 129–132) cite the study *Revisiting the American white working class 1985 to 2000*, in which a researcher in 2000 reinterviewed participants from a study she had conducted in 1985. In this case, the researcher believed that her prior relationship with her subjects was integral to how they responded in the later round of interviews; a different researcher would have come away from the same subjects with different results. Making data gathered in social science fieldwork available for reuse must also navigate the complex and fraught terrain of the privacy, consent and trust of its subjects. Although McLeod and Thomson are particularly interested in the reuse of qualitative data gathered by social science researchers, their discussion of reusing datasets is relevant to the data reuse discussion more broadly. They have observed that "the imprint of the original researcher's identity permeates the archive – in notes, selection of materials and so forth" (McLeod & Thomson, 2009, p. 136). This can be applied to any research dataset as the priorities of the original research project will inform the data collected and eventually affect the applications for which it is suitable. Even in the context of quantitative research, contextual factors such as the precision of equipment used to take measurements or the sample size should also be considered when approaching a dataset.

3. For the purposes of this paper, I am adopting the definition of De Mauro, Greco and Grimaldi (2016, p. 131) that big data is "the Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."

In addressing these problems, McLeod and Thomson turn to methodologies from history, which have been developed in deep engagement with how to reanalyse sources—both primary and secondary. Historians "need to reconstruct and reimagine […] the context and time in which the material was produced – these are part of the creative and intellectual challenge, not regarded as obstacles or reasons not to undertake the work" (McLeod et al., 2009, p. 138). Researchers working with datasets—especially those they have not themselves compiled—must recognise that a dataset is the product of a particular moment in time; each researcher approach it anew with their own questions and preoccupations. Source criticism is a tool of historians that has application in research far beyond historiography. I do not wish to suggest that researchers from the natural sciences have not considered these issues of data reuse. A 2014 note on secondary analysis of existing data argues that "researchers must have a comprehensive understanding of the strengths and weaknesses of the [given] dataset" (Cheng & Phillips, 2014, p. 373). However, as the growing availability of both data and analytic tools has created opportunities for researchers to venture further from their core areas of expertise, critical attention to data becomes more important. Researchers working with datasets from outside their core discipline run an increased risk of misunderstanding a critical element of a dataset that would be plain to an expert in the field and its methodologies. Oakden-Rayner, as both a radiologist and a machine learning researcher, was uniquely placed to identify the limitations of the ChestX-ray14 dataset. It is easy to imagine, however, computer scientists who lack his ability to cross-check results with their own interpretation of the X-ray images, relying instead on the labels and training of a flawed model. Scientific studies should acknowledge and explore their limitations and source criticism must become a key component of the researcher's toolbox, regardless of their discipline.

COMPUTATIONAL TOOLS

If the Humanities and Social Sciences offer nuanced methodologies for approaching data, they have sometimes been less confident in approaching the computational tools used to process it. As Fiormonte (2012, p. 60) has argued, however, technology is a cultural artefact and, like any other cultural artefact, it "is subject to the influence of its environment, culture, and the social habits of the individual and groups that devise and make use of [it]." Fiormonte (2012, p. 62) goes on to argue that, by focusing on digital archives and data analysis, the digital humanities have neglected to engage sufficiently the capacities of computation to affect research processes. In the years since Fiormonte's article appeared, lively critiques of big data and machine learning have emerged from the social and political sciences as challenges to the myth that computational techniques transcend human foibles.

Opening up code bases, however, is not in itself going to guarantee that technology will become comprehensible and accountable to humans. Instead, digital journalist Diakopoulos (2015, p. 400) suggests a means of interrogating algorithms based on the four main operations that he sees them performing: prioritisation; classification; association; and filtering. He stresses that algorithms are the products of human design—whereby active choices are made by their developers concerning criteria, training data, and semantics—and that algorithmic accountability "must therefore consider algorithms as objects of human creation and take into account intent […] that may have influenced their design as well as the agency of human actors in interpreting the output" (Diakopoulos, 2015, p. 402). He further suggests a range of criteria that may be useful for interrogating the 'transparency' of algorithms, including the data used, the rates of false positives and false negatives, and the criteria used to design the algorithm (Diakopoulos, 2015, p. 411). This approach recognises the human design decisions that are applied in both the development and the implementation of the algorithm, satisfying Fiormonte's (2012, p. 60) demand that code be approached as a cultural artefact and interrogated as such.

This faith that machine learning can extract meaningful information from datasets without human supervision and guidance has led to some embarrassing errors. A well-known example is that of Google Flu Trends. In 2008 Google announced that, by monitoring search terms, it could predict outbreaks of the flu, offering a valuable tool for public health planning (Kennedy & Lazer, 2015, para. 2). This tool, however, failed conspicuously in 2013 when predictions were out by 140% (Kennedy et al., 2015, para. 3). In a critique of insufficiently rigorous data mining studies, Smith (2019, para. 5) has argued that rather than allowing the model to identify search terms that correlated with flu outbreaks, the team behind Google Flu Trends would have done better to select flu-related terms to train their model. By using unsupervised learning, the model attached significance to search terms such as 'high school basketball,' which has some correlation with the American winter and flu season but is in fact unrelated to the flu as such (Kennedy et al., 2015, para. 6). In his article Smith (2019) argues that studies need to be rigorously designed, rather than simply spelunking through data. He has stated that "good research begins with a clear idea of what one is looking for and expects to find. Data mining just looks for patterns and inevitably finds some" (Smith, 2019, para. 14).

Researchers working with datasets from outside their core discipline run an increased risk of misunderstanding a critical element of a dataset that would be plain to an expert in the field and its methodologies

MAKING THE WORLD THROUGH DATA

Despite its known limitations, machine learning is being used to automate a wide variety of decision-making with sometimes terrible results. This is especially the case for individuals who find themselves unable to secure employment or access healthcare due to being categorised as undesirable. In recent years a growing body of literature dedicated to investigating these injustices has emerged. With names like *Weapons of Math Destruction* (O'Neil, 2016), *Automating Inequality* (Eubanks, 2017), *Technically Wrong* (Wachter-Boettcher, 2017) and *Algorithms of Oppression (*Noble, 2018), this literature makes the case that technologists have been insufficiently thoughtful of the consequences of their work, especially on already marginalised groups. These criticisms have not gone unheeded. Initiatives such as Google's publication of its *Objectives for AI Applications* (Pichai, 2018) seek to reassure the public that Google is actively engaged with questions of safety, accountability and public interest in its research. This interest in understanding, critiquing and ultimately holding algorithms and their creators accountable for the outcomes of their work has, however, encountered limitations in the usefulness of normative ethics and notions of transparency.

More profound challenges to the basis of data-intensive research are summarised in Mauthner's 2018 article *Toward a Posthumanist Ethics of Qualitative Research in a Big Data Era.* As discussed above, Mauthner (2018, p. 21) critiques 'big data positivism'; the assumption that a sufficiently large and complex dataset must, by its nature, approach some objective reality. Her criticism is not, however, limited to the problem of the fitness of any dataset to answer research questions. Rather, Mauthner (2018, p. 3) draws on post-foundational thinking to reject the assumption—rooted in the Enlightenment—that an objective reality, accessible to human researchers via sufficiently rigorous methods, exists at all. In contrast, she argues that research *makes*, rather than simply *uncovers*, the world:

> The world is not composed of pre-existing and already-formed entities awaiting discovery by human knowers, whose ethical responsibility is to ensure that these entities are accurately represented and in a way that avoids harm. Rather, knowledge practices are understood to play a constitutive part in bringing their objects of study into existence. (Mauthner, 2018, p. 3)

The idea of ethical research here moves from the concerns of normative ethics—that is, how research subjects should be protected from harm while participating in research—to questions of the world that is being created by research. That is, there is no distinction, for example, between 'finding out about the world' and 'ensuring that no harm is done in the process.' Rather, Mauthner (2018, p. 12) argues that "there is ethical duty and responsibility in knowledge/world-making itself. Knowledge production is an inherently ethical matter." Described in these terms, Mauthner's theory can sound too abstract to be of much use to practitioners of data analytics, where it in fact poses a deep challenge to our ideas of the uses to which data about humans can be applied.

One of the most significant criticisms of automated decision-making based on large datasets is that this process disadvantages already marginalised individuals due to systemic issues such as over-policing of some populations that affect the nature of the dataset itself. A striking example is that of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), software used to assist in decision-making in the US justice system. The model routinely scores coloured defendants as more likely to reoffend than their white counterparts, leading to higher bail amounts and longer custodial sentences (Angwin, Larson, Mattu, & Kirchner, 2016; Wachter-Boettcher, 2017, pp. 119–121). Critics have argued that coloured populations both receive harsher penalties than whites and are over-policed as a group, meaning that the training datasets reflect the treatment of people of colour by the police and courts. An original investigation into the COMPAS software by Pro Publica found that the problem was not merely racial: of all people flagged by the software as likely to commit violent crimes, only 20% went on to do so (Angwin et al., 2016, para. 13). When all offences, including misdemeanours, were taken into account, only 61% of those flagged as likely to reoffend were arrested again within two years (Angwin et al., 2016, para. 14). Post-foundational approaches invite us to consider this problem in a new ethical light. Rather than assuming that 'likeliness to reoffend' is a characteristic inherent to a person, and the problem is correctly calculating its probability, we are forced to ask on what basis any human or algorithm creates such a category and assigns it to others. Profound questions about the nature and purpose of the justice system are opened up once we entertain the idea that 'likely to commit a crime' is a label created and applied to individuals, rather than the discovery of an objective truth. The COMPAS algorithm is by no means an isolated example of the sorts of information that the miners of big data claim to be able to uncover, but arguably create, from 'likely to commit violent crime' to 'likely to buy movie tickets.' Forms of artificial intelligence have been created that claim to be able to identify the ideal employee, borrower, and student. Accepting that these categories are called into being to serve human purposes allows us to challenge them at a deeper level than examination into data and algorithms alone generally permits. Here we are able to ask: who is, and who is not, served by their creation?

FROM ETHICS TO JUSTICE

As the growing literature makes clear, there is mounting concern about ethics in the field of AI. Many companies that undertake research and development in these fields are starting to appoint their own data ethics advisors and signal their interest in ethics. In 2016, for instance, Accenture Consulting's digital labs published two discussion papers on data and algorithmic ethics (Lynch et al., 2016; Turner, David, & Wulfsohn, 2016). Then in 2018, Microsoft launched its AI for Good program (Smith, 2018), the same year that Google published AI Principles (Pichai, 2018). While the particularly public increase in interest in the ethics of data practices is welcome, this is the beginning and not the end of this journey. Expressing the intention to behave 'ethically' does not in itself offer much insight into one's intentions, but rather invites uncritical approval. We need a language with which to interrogate the use and impact of assemblages of data and algorithms, especially when these applications occur outside of the bounds of traditional research institutions.

Traditional research ethics frameworks struggle to accommodate research done on big data, as analysis may be done in contexts divorced from the creation and collection of the data in question. The Australian *National Statement on Ethical Conduct in Human Research* and accompanying *Australian Code for the Responsible Conduct of Research* (NHMRC, 2018a & 2018b) acknowledge that data governance is now part of research. Their principles-based approach, however, leaves researchers to interpret for themselves how to approach data that derives from human activities, such as social media posts, web analytics, mobility data derived from wifi usage and the plethora of other data types that have become available via an increasingly digital world. The models of responsibility between researcher and subject struggle to accommodate research where the two are increasingly removed from each other as individuals and the subject is known only via social media handles and the identifiers of devices.

Practices in collecting data and organising our world based on the results of analysis mean that the question of ethics cannot be ignored, even as research ethics frameworks are challenged by the possibilities of big data. If our world is in some measure created through data, it is incumbent on us to ask what sort of a world we are building. Here, I am turning to the question of justice, specifically the analysis of social justice offered by Fraser (1998) in her essay *Social Justice in the Age of Identity Politics: Redistribution, Recognition, and Participation*. In it, Fraser (1998, p. 1) sets out to articulate two types of claims for social justice: redistributive—which focuses on just distribution of goods—and recognition—which considers who is represented, and how "assimilation to majority or cultural norms is no longer the price of equal respect." I have chosen this as a starting point as it calls specific attention to the tensions and inequities that exist in human societies, that are replicated into the datasets and codebases generated by those societies. By way of example, Noble's (2018) work investigates how online discoverability has been created in ways that systematically discriminate against black women. Social theory, which seeks to understand the relations between "class and status, economy and culture" (Fraser, 1998, p. 1) offers a lens through which to understand the nature of the harm perpetrated against black women, in this case, and how it operates.

While Fraser wrote this paper in 1998 when thoughts of racist algorithms were far away, her framing of social justice and her insistence that economic and cultural factors cannot be separated from each other apply remarkably well in addressing the current realities of these issues. First, Fraser (1998, p. 5) raises the notion of "parity of participation." To deny some individuals and groups the ability to participate in society on equal terms because of certain characteristics is a failure of justice. Furthermore, treating this as an issue of social justice rather than individual relations reframes the problem from being one of relations between individuals to a problem of institutional patterns. Systemic inequality designates some members of society as 'worth less' than others and denies them "the status of [a] full partner in social interaction[s]" (Fraser, 1998, p. 3). This framing allows us to understand an algorithm that returns racist search results is perpetrating an injustice. Beyond offending individual black users, it has denied this population, on a collective basis, the ability to participate equally in online social relations. Furthermore, this framing elevates individuals who do not conform to the ideal user envisaged by product designers from being 'edge cases' to victims of injustice, making their claims impossible to ignore. Secondly, Fraser (1998, p. 8) argues for what she calls "perspectival dualism" whereby economic and culture factors cannot be separated from each other but are rather treated as two aspects of any domain. She points to the way gender norms (cultural) are intertwined with labour practices (economic), which combine to form ideas of men's and women's work (Fraser, 1998, p. 10). Again, this framing allows researchers to see questions of representation and access as part of a much bigger pattern, elevating it beyond the struggles of individuals.

Fraser's approach, which considers both the importance of allowing equitable access to resources and gives space to the complexities of individual identities and needs, provides a useful tool for assessing the workings of algorithmic assemblages. Calls for algorithmic ethics have often focussed heavily on the code itself. As argued above, while it is important to recognise that computer programs are the result of human decision-making, the opening up of code bases is not enough to ensure accountability (see for instance Ananny & Crawford, 2018 and Kemper & Kolkman, 2018 on the limits of 'transparency' as an accountability mechanism). Rather, as Ananny (2016, p. 109) argues, a framework for algorithmic ethics needs to focus on how the assemblage acts. Fraser's (1998) model of social justice—in considering the dual axes of economic and cultural conditions that affect how a group or individual is able to participate in society—offers such a framework.

CONCLUSION

Research methods from the Humanities, Arts and Social Sciences have much to offer the emergent disciplines that utilise big data and analytical algorithms. Perhaps the most important concept is the basic premise of post-foundationalism itself—that there is no objective reality to be captured in a dataset or reflected in a model. Rather, all assemblages of data and code are the product of human selection and decision-making. In choosing to work with these datasets, we must pay attention to their specific contexts rather than merely attempting to move beyond or behind them. A striking example of the extent to which cultural associations are embedded in data comes from studies of word embeddings—a means of textual analysis that studies co-occurrences of words in textual corpora (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). The authors describe the results when a computer program trained on a corpus of Google news articles was asked to complete analogy pairs: "the [...] system will answer "man is to computer programmer as woman is to x" with x=homemaker. Similarly, it outputs that a father is to a doctor as a mother is to a nurse" (Bolukbasi et al., 2016, p. 3). The authors express dismay at this outcome, characterising the first example as "offensive" and saying "one might have hoped that the Google News embedding would exhibit little gender bias because many of its authors are professional journalists" (Bolukbasi et al., 2016, p. 3). The issue here is not that the individual journalists whose works make up the corpus are 'biased', but rather that labour and gender are deeply entwined in our society and these conditions are reflected in the dataset. As Fraser (1998, p. 8) argues, "nominally economic matters usually affect not only the economic position but also the status and identities of social actors." Once the relationship between gender and occupation was elucidated, the researchers experimented with changing this weighting in an attempt to produce a dataset devoid of 'gender bias.' While the quest to create 'unbiased data' is doomed to fail—as it is predicated on the notion that a pure and objective reality exists and can be accessed—sensitivity to the interplay of power, identity, and culture in the creation of any dataset will allow it to be used in more critical and nuanced ways.

Sensitivity to the contexts in which datasets are assembled and the preoccupations of the researchers who analyse and reanalyse them is essential to working effectively with any data, including so-called big data. Additionally, researchers must recognise that computational tools are also the product of human decision-making at a particular moment in time; faith in the ability of machine learning algorithms to uncover patterns does not replace research design. Such historiographic approaches to data and code offer a means of working with both in ways that will be respectful of their limitations, engaged with their strengths, and sensitive to the ethical implications of knowledge-making.

REFERENCES:

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. doi.org/10.1177/1461444816676645

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values*, 41(1), 93-117. doi.org/10.1177/0162243915606523

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bolukbasi, T., Chang, K-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Retrieved from http://arxiv.org/abs/1607.06520

Broad, E. (2018). *Made by Humans: The AI Condition*. Carlton, Australia: Melbourne University Press.

Cheng, H. G., & Phillips, M. R. (2014). Secondary analysis of existing data: opportunities and implementation. *Shanghai Archives of Psychiatry*, 26(6), 371-375.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). *A formal definition of Big Data based on its essential features. Library Review*, 65(3), 122-135. doi.org/10.1108/LR-06-2015-0061

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398-415. doi.org/10.1080/21670811.2014.976411

Eubanks, V. (2017). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (First Edition). New York, USA: St. Martin's Press.

Fiormonte, D. (2012). Towards a Cultural Critique of the Digital Humanities. *Historical Social Research*, 37(3), 59-76.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2), 179-188. doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fraser, N. (1998). Social justice in the age of identity politics: redistribution, recognition, participation. *WZB Discussion Paper,* 1, 98-108.

Gainsford, P. (2018). The citation problem. Retrieved from http://kiwihellenist.blogspot.com/2018/09/the-citation-problem.html

Kemper, J., & Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 1-16. doi.org/10.1080/1369118X.2018.1477967

Kennedy, R., & Lazer, D. (2015). What We Can Learn From the Epic Failure of Google Flu Trends. Retrieved from https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Lynch, H., Bartley, R., Metcalf, J., Petroni, MJ., Ahuja, A., & David, S. L. (2016). Building digital trust: The role of data ethics in the digital age. Retrieved from https://www.accenture.com/_acnmedia/PDF-22/Accenture-Data-Ethics-POV-WEB.pdf

Mauthner, N. S., & Parry, O. (2013). Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology: A Journal of Knowledge, Culture and Policy*, 27(1), 47-67. doi.org/10.1080/02691728.2012.760663

Mauthner, N. S., (2018). Toward a Posthumanist Ethics of Qualitative Research in a Big Data Era. *American Behavioral Scientist*, 63(6), 669-698. doi.org/10.1177/0002764218792701

McLeod, J., & Thomson, R. (2009). *Researching Social Change: Qualitative Approaches*. London, UK: SAGE.

Miranda, P. J., Baptista, M. S., & de Souza Pinto, S. E. (2018). The Odyssey's mythological network. *PLOS ONE*, 13(7). doi.org/10.1371/journal.pone.0200703

National Health and Medical Research Council. (2018a). Australian Code for Responsible Conduct of Research, 2018. Retrieved from https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018

National Health and Medical Research Council. (2018b). National Statement on Ethical Conduct in Human Research (2007) - Updated 2018. Retrieved from https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, USA: New York University Press.

Oakden-Rayner, L. (2017). Exploring the ChestXray14 dataset: problems. Retrieved from https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (First Edition). New York, USA: Crown.

Pichai, S. (2018). AI at Google: our principles. Retrieved from https://www.blog.google/technology/ai/ai-principles/

Smith, B. (2018). Using AI to help save lives. Retrieved from https://blogs.microsoft.com/on-the-issues/2018/09/24/using-ai-to-help-save-lives/

Smith, G. (2019). The Exaggerated Promise of So-Called Unbiased Data Mining. Retrieved from https://www.wired.com/story/the-exaggerated-promise-of-data-mining/

Turner, C., David, S. L., & Wulfsohn, G. (2016). Ethical Algorithms for "sense and respond" systems. Retrieved from https://www.accenture.com/_acnmedia/PDF-27/Accenture-Ethical-Algorithms-Digital.pdf

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech* (First Edition). New York, USA: W.W. Norton & Company.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Retrieved from http://arxiv.org/abs/1705.02315

any system of data collection and use is always bound to a framework—either it cultivates inclusion or, failing that, it systematically works to exclude low-income, minority or underserved communities from, for example, access to society's broader benefits

BEYOND BIGNESS: CAN BIG DATA HAVE AN ETHICAL FUTURE?

By Tyne Daile Sumner

During occasions when new industries and new technologies are developed, the physical and physiological details usually taken as given can become a matter of concern with consequent clarification of the assumptions and conceptions we have of what individuals are.

—Erving Goffman (1969, p. 4)

ABSTRACT

Questions of privacy and security have long been associated with the collection and use of big data. Increasingly, however, critics have come to associate big data with concepts such as fairness, accountability and transparency. As the scale and complexity of data continues to expand, scholars are being called upon to offer up new ways of tackling the complex ethical debates at the centre of big data. Research presented here covers a range of issues that cut across social, political and technological applications to consider the effects of large-scale data in the twenty-first century. By asking 'Who is missing?' in big data, we can, by extension, consider questions such as: 'What are the ethical debates that matter?' and 'What still needs to be considered beyond the basic codes and protocols for the governance of data?.' Moreover, one of the first necessary steps in rethinking big data's ethical future is a reconceptualisation of the very notion of 'bigness.'

INTRODUCTION

The collection and use of data is unquestionably one of the central preoccupations of our time. More recently, big data has come to dominate social and political fields as diverse as medicine, business, advertising, social media and the news. Described by Mayer-Schönberger and Cukier (2013, p. 6) as "things one can do at a large scale that cannot be done at a smaller one," big data has also begun to have profound effects on the ways we see ourselves as citizens of the world, such that it now clearly shapes our subjectivities (Schroeder, 2018, p. 127).[4] Labelled by some scholars as 'datafication,' this worrying trend can be described as the "quantification of social interaction and their transformation into digital data" (Richiterich, 2018, p. 1). Via this formulation, not only are our tastes, preferences and choices affected by data, the way in which we relate to other people and articulate ourselves is intimately connected to big data's influence. As renowned Information Studies scholar Christine Borgman reminds us, however, "big data is not necessarily better data." She notes the way in which, the farther the observer is from the point of origin, "the more difficult it can be to determine what those observations mean - how they were collected; how they were handled, reduced, and transformed; and with what assumptions and what purposes in mind" (2015, p. xvii).

Comments such as this should alert us to the fact that any system of data collection and use is always bound to a framework—either it cultivates inclusion or, failing that, it systematically works to exclude low-income, minority or underserved communities from, for example, access to society's broader benefits. Moreover, as Eubanks (2018, p. 7) points out in the revelatory *Automating Inequality: How high-tech Tools Profile, Police and Punish the Poor:*

[M]arginalized groups face higher levels of data collection when they access public benefits, walk through highly policed neighbourhoods, enter the health-care system, or cross national borders. That data acts to reinforce their marginality when it is used to target them for suspicion and extra scrutiny. (see also Browne, 2015; Lyon, 2003; Mann & Daly, 2018; Noble, 2018; Gangadharan, 2012; and Sandvig, Hamilton, Karahalios & Langbort, 2016).

While some forms of data-driven inequality arise via the removal of certain people or voices from a collective, others employ digital surveillance technologies to over-monitor particular people or groups of people, usually to damaging effect (Ferguson, 2017). This article gathers a range of recent reflections in the field, alongside a short history of the concept of privacy, to suggest new ways of tackling the complex ethical debates at the center of big data. The relevance of concepts such as justice, empathy, agency, ownership, privacy, subjectivity and identification are also of considerable importance to the disscusion that follows.

4. Schroeder (2018, p. 127) offers another dimension to this definition, noting: "'Big data' can be defined as research that represents a step change in the scale and scope of knowledge about a given phenomenon."

BIG DATA: HOW DID WE GET HERE?

Of the three 'v's' that have come to define big data—volume, variety and velocity—it is perhaps volume that has attracted the most attention.[5] Most information that was formerly stored in wallets or filing cabinets is now digital and growing at an accelerating pace. This is a technological development that has afforded unprecedented data access to more people than ever before. Yet while there is an almost global understanding that personal data should be protected, an individual's private information is nevertheless susceptible to the same exploitative systems that have historically infiltrated other social phenomena: trade, economics, politics, education, climate and so on. Of course, all of these things are inextricably connected to big data, yet somehow the narrative has prevailed that the mass collection of an individual's private information sits outside the realm of the day-to-day functioning of a society.

One way out of this paradox is to foreground the fact that big data refers not only to data, as such. Rather, it encompasses the behaviours, practices, networks, infrastructures and politics that influence and are influenced by its manifestations. Understanding these overlaps is one way of understanding big data as a set of "emerging technology" practices since it also encompasses "digitally enabled developments in data collection, analysis, and utilisation" (Richterich, 2018, p. 23). Moreover, one of the reasons why big data has been increasingly tied to broad debates about human rights, autonomy, transparency, privacy, security and self-responsibility is because it fundamentally challenges pre-existing moral and ethical norms. It does this by advancing the cumulative knowledge of data-collecting organisations, thereby also advancing the power gained over individuals and groups. A now widely-understood offshoot of this trend is the "application of big data knowledge in shaping media uses," which thereby has dramatic effect on the social implications of such media usage (Schroeder, 2018, p. 127). Big data poses a challenge also to pre-existing ethical frameworks with regards to consent, insofar as corporate data economies have succeeded in organising big data's alleged lack of bias towards commercial gain. It is for this reason that any intellectual inquiry into big data requires a more robust methodology than simply probing the data itself; it also necessitates an interrogation of the knowledges and power structures that underpin its collection and usage in the first instance. In order to address this issue, many scholars have come to see big data as an overarching framework for understanding the contemporary technological landscape. In *We Are Data: Algorithms and the Making of our Digital Selves,* Cheney-Lippold (2017, p. 4) notes, for example, the ways in which our "daily activites are mediated with software" such that the "resulting aggregation of our lives' data founds the discursive terrain of our digital environments." Similarly, others note the way in which big data functions as a crucial "sense-making" resource in the digital era (Andrejevic, 2014, p. 1675).

While these are useful ways of thinking about big data's ubiquity and dominance, these datasets nevertheless lend themselves to problematic misinterpretation (Harford, 2014). One way in which the pervasiveness of big data is prone to being misunderstood is in the assumption that size and scale somehow equate to lack of bias. The common coinage 'digital positivism' goes some way towards explaining this assumption insofar as it encapsulates a range of theories that assert that data—in ways similar to the physical world—operates according to general or universal laws (Mosco, 2015 & 2016). We need look no further, however, to the now common example of big data policing—euphemistically known as 'predictive reasonable suspicion'—to know with certainty that more data does not necessarily correlate to more ethical data systems. With more specific information, police officers may now be afforded a stronger predictive sense that they are in fact observing a criminal act. As Ferguson (2015, p. 331) points out, however, the "next phase will use existing predictive analytics to target suspects without any firsthand observation of criminal activity, relying instead on the accumulation of various data points." The underlying problem with this formulation is that the very data used for predictive purposes contains built-in sociodemographic bias. Or, to borrow from Ferguson (2015, p. 331) again, "this new reality simultaneously undermines the protection that reasonable suspicion provides against police stops and potentially transforms reasonable suspicion into a means of justifying those same stops."

One way in which big data holders and organisations have attempted to defend against this reality is by asserting the relevance of informed consent. While there exists an abundance of material on 'best practices' for informed consent in relation to data—especially for enabling the reuse of research data beyond the purpose for which it was collected—the fuzzy ethics surrounding informed consent cannot be ignored (Koops, 2014; Gellert & Gutwirth, 2013; and Parsons, 2015). Part of this problem is with regards to transparency, while there is also the complicating factor of limited citizenry knowledge of how big data actually operates. "Often, when confronted with the potential of using personal data," Matzner (2014, p. 96) comments, "people react surprised and affected suggesting that they would not have consented to this use of their data if they had been informed about the possible consequences." This is further complicated by the fact that even in situations where it is possible to acquire comprehensive information about the nature, collection and use of data, the complexity of the process and effort involved poses ethical problems in itself (van der Ploeg, 2007, p. 49). As Matzner (2014, p. 96) goes on to assert, "it is questionable if such an effort can reasonably be required by everybody or whether this establishes new inequalities in terms of knowledge and skills necessary to use a service or technology." Thus, the bigness of big data in turn creates a situation whereby processes of comprehension or discernment itself can generate new and potentially unequal power structures.

5. Note that some scholars have sought to add a fourth 'v'—veracity—to this paradigm, in order to draw attention to questions of reliability around certain data usage.

Recent personal data breaches by Facebook, for example, have drawn worldwide attention to the problematic and slippery frameworks that underpin the data collection processes for many large corporations. Following widespread uproar over the Cambridge Analytica data leak, Facebook now restricts developer access to user data. It is a move that Schroepfer (2018, para. 1), Facebook's chief technology officer, described as a change that will "better protect people's information while still enabling developers to create useful experiences." Yet while protecting user data from potentially exploitative developers and data-hungry apps might seem like an ethical response to Facebook's privacy problems, the reverse is in fact the case. By restricting access to various Application Programming Interfaces (APIs) and thereby reducing data transparency, Facebook has instead successfully leveraged a massive security data breach to make it harder for outside groups (reseachers, for example) to gain insight into its algorithmic objectives. While this recent protective measure can work on the one hand to block access by developers to users' religious preferences, it also operates on the other to prevent research into Facebook's targeted advertising processes.

As recent events have revealed, these data organising principles are often closely connected to the overarching ideology behind large businesses or corporations. The collection of big data and its associated algorithms reflect broad conceptions of power that are akin to Foucault's (1975) influential model in which power is less a force exerted on individuals and rather a dynamic deeply embedded within societies at large. Facebook's data breach is just one example that exposes the complex entanglements between consent, ethics and corporate big data practices; as we are becoming increasingly aware, there will be more of the same to come (Kramer, Guillory, & Hancock, 2014).

Moroever, the well-known Facebook case highlights a common grey area when it comes to informed consent in the context of big data (PrivazyPlan, 2018).[6] Ultimately, the approval by users of social media privacy policies does not seem sufficient enough grounds for corporations to justify the use of personal data for any variety of commercial or research purposes; rather there should be some clear limitations put in place to further protect these practices (Rothstein & Shoben, 2013; Ioannidis, 2013). In unpacking these views and others, there is also need to consider the role of algorithmic bias; the way it reflects not just the techno-corporate contexts in which the majority of big data is created, but also the political and educational frameworks and organisations through which this data moves and is governed. While the open data movement promotes the accessibility of data as a public good, not all data is created equal nor do all citizens have equal access to it. As Richterich (2018, p. 40) has usefully written in relation to this point, "the 'big data divide' implies power/knowledge conditions that systematically exclude individuals from access to data which would allow them to assess the data generated by corporations, the conditions under which this is done, and how this information is used." This is indeed a key barrier to effectively tackling algorithmic bias at a deeper level regarding unequal access to data in the first place (Powles & Nissenbaum, 2018).[7]

PRIVACY: IS THERE ANY LEFT?

Placing some of the more pressing concerns associated with big data in the context of privacy's long-term erosion might shed some light on whether citizens' concerns have historically had any effect on the organising principles of those who collect and use individuals' data. Afterall, widespread concern over privacy is in no way unique to the current moment. In their seminal essay *The Right to Privacy*, Warren and Brandeis (1890, p. 205) pronounced that the right to privacy was based on a principle of "inviolate personality," thus laying the foundation for the modern understanding of privacy as control over one's personal information. Later, Westin (1967, p. 7) defined privacy as the "claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others." By the second half of the twentieth century, anxiety around loss of privacy was pervasive. This anxiety was generated in part by new visual and audio technologies, as well as changes in constitutional privacy laws; it was largely due, however, to the intensification of surveillance activity in the early years of the postwar period. A surge of writing emerged in response to such trends, which aimed to not only highlight the large-scale collection of individuals' information by government and corporate bodies, but also the need for a collective ideological resistance to such trends. Dash's 1959 publication The Eavesdroppers opened the gate for an outpouring of texts that examined privacy through an unprecedented sociological lens, as a topic requiring urgent critical attention. Subsequent texts such as Ernst's *Privacy: The Right to Be Let Alone* (1962), Brenton's *The Privacy Invaders* (1964), Westin's *Privacy and Freedom* (1967), and Smith's *Privacy: How to Protect What's Left of It* (1979) represent a snapshot of the period's intense focus on the problems associated with the rapid erosion of personal privacy. Collectively, these texts signal that by the end of the twentieth century, citizens the world over were beginning to accept that the boundaries between their private and public selves were no longer secure. An argument common to many of these publications is that privacy as a concept is, by its very nature, linked to notions of personhood and self-identity (in Kulhari, 2018). The right to privacy—also known as informational self determination—is an "important facet of the right of personality, which guarantees every individual the possibility to develop her own personality" (Kulhari, 2018, p. 28). While current practices of handing over personal data are frequently indirect and ancillary—one

6. While I am using the concept of 'informed consent' in the context of this discussion, it is worth noting that consent is not the only legal basis for processing information (although it depends on jurisdiction). For example, the EU General Data Protection Regulation (GDPR) sets forth six conditions for the lawfulness of processing data: "consent; for the performance of a contract; for compliance with a legal obligation; to protect the vital interests of the data subject or of another natural person; for the performance of a task carried out in the public interest; and for the purposes of the legitimate interests pursued by the controller or by a third party" (PrivazyPlan, 2018).
7. Powles and Nissenbaum (2018, para. 7) consider the extent to which trying to 'fix' AI and algorithmic bias actually distracts from the more urgent questions about the underlying technology used in these systems, as well as the unequal power structures that underpin the data that comprises them in the first place.

example being the divulging of specific tastes and preferences via Social Media platforms—earlier narratives were ones in which a person's subjectivity was not yet modulated or externalised via big data. Despite these changes, the protection of personal data within democratic societies today still tends to be considered an extension of the right to privacy, despite arguments that they are distinct. Scholars who justify the inextricable connection between privacy and data protection frequently foreground claims for the right to data protection being characterised by strong links to the right to privacy (Gonzalez Fuster, 2014). Those who argue that privacy and data protection rights are substantially distinct, frequently invoke the scope and size of each respective category, arguing that although the two often overlap, there are instances of data processing that have nothing to do with personal privacy (Gellert et al., 2013, p. 525).

Increasingly, privacy has less to do with the ways in which individuals choose to disclose personal information, and more to do with the ways in which they interact either directly or indirectly with a wide array of social, political and cultural phenomena. Because big data operates most successfully within overlapping realms of public and private, these practices seem to somehow elude many prewar classifications of subjectivity. Prior to the rise of mass electronic surveillance during the pre-Internet era of the mid-to-late twentieth century, a person's sense of who they were, together with what status their personal data occupied, was relatively unimpeded by technological frameworks. To put this another way, an individual's personal data previously existed in a comprehensible form. As this scenario changed, so too did our collective understandings and expectations of how much autonomy over personal data an individual could and should have. Today, the rise of big data has put the very idea of an individual's self-hood and autonomy under direct threat. Big data critics have reflected this by emphasising the lack of control, knowledge and agency that individuals have over the ways in which their personal information is being collected in relationship to the use of online services (Tene & Polonetsky, 2012). There is also concern related to the insistence of service providers that a user's personal data remains anonymous; an assurance which critics have come to see as almost impossible. Richterich (2018, p. 38) has summarised this concern, where she states that:

Big data enforce an increased, though neither necessarily deliberate nor conscious transparency of online users/consumers. The full extent of this transparency is only visible to those actors controlling the main data collecting platforms or gaining external access to these. What is ultimately collected here, are vast amounts of personal information concerning individuals' preferences, attitudes, moods, physical features and […] health status and health-relevant behaviour.

Questions surrounding data collecting platforms often highlight related privacy violations when it comes to the reidentification of an individual's data, in particular, data stored in either a private or public health record. Public health data exists at the crossroads of several big data tensions, offering useful examples for thinking through the interdependencies of big data practices with forms of health surveillance, scientific research and ethics. Indeed, when we begin to think about the relationship between privacy and processes of data reidentification, several questions come to mind. In the first instance, what does it *mean* for a platform to have 'stored' personal data? Second, what are the implications of frameworks that seek to re-identify data that has previously been de-identified? And finally, do data collecting organisations have an ethical and/or moral responsibility to notify individuals whose data has been intentionally or accidentally leaked? An ostensible 'quick fix' to some of these questions has been the suggestion that banning deidentification practices at the outset of a data collection process would prevent the potential for any subsequent breach. Such drastic measures undoubtedly generate further problems by preventing people who are trying to gain transparency around particular data systems from in-turn interrogating them. Thus, criminalising the reidentification of an individual's data, in any context, might provide enhanced citizen confidence and surety in the short term, but does not actively make the system more ethical in the long term.

ETHICS: WHO IS MISSING AND HOW?

By virtue of its 'bigness,' big data ultimately fosters a culture of what might be called data noise or data saturation. Indeed, critics have recognised this trend as early as the mid-twentieth century, albeit as a product of technologies such as television—now viewed as almost benign in comparison to something like Facebook's election-fixing algorithmic capabilities. As early as the 1950s in America, for example, new forms of media, combined with the corporatisation of modes of communication, were creating a culture in which "public discourse belong[ed] entirely to the mass media, particularly electronic media, [to] include only the voices of those who could penetrate or manipulate a genre of discourse that thrive[d] on overcommunication" (Doreski, 1999, p. 75). Within this arrangement, those who have the power to shape and control the flow of data also have the capacity to prioritise particular narratives. While the machinations and effects of online news and advertising algorithms are the source of much recent critical attention, the social and political implications of big data's tendency to enact processes of exclusion and, conversely, over-inclusion still requires more consideration. The Data Justice Lab at Cardiff Univeristy in the UK represents a key organised approach to articulating these implications, via its development of a research agenda focused entirley on examining the complex relationship between datafication and social justice. The lab maintains, for example, a Data Harm Record, which runs a continual log of problems associated with

automated and algorithmic systems reported from across the globe (Redden, 2018, para. 4). The record divides the broad concept of 'data harms' into eight useful cateogries: commercial uses of data (potentials for exploitation); discrimination; loss of privacy; identity theft, blackmail, reputational damage and/or distress; physical injury; political uses of data, political manipulation and social harm; Government uses of data (data errors); and harms due to algorithm/machine bias (Redden & Brand, 2018).[8] Across all of these groupings, what stands out is a common thread of power imbalance; that is, between those who collect and hold data, and those whom the data is ostensibly about. The new forms of categorisation enabled by the collection of big data are often created without our knowledge and are "based on criteria that do not necessarily correspond to lived experience" (Dencik, Hintz, Redden, & Warne, 2017, p. 734). Ultimately, a truly ethical approach to big data needs to move beyond mere analysis of what particular algorithms achieve—via the collection and manipulation of personal information—towards a more complex interrogation of what gets lost amidst the noise.

The obvious answer to this problem is to build greater equity into infrastructure systems. But how? One possible way is to transform the organising principles of data ethics by moving them away from a consideration of the 'average person' and rather highlighting those who are the most vulnerable and marginalised. It can be argued that this motion foregrounds the relevance of *justice* within big data practices by focusing on the deconstruction of power asymmetries and marginalisation (Taylor, 2017, p. 20; Johnson, 2014; Heeks & Renken, 2018). The problem with this approach, however, is that any form of "data-centric rationality" is always tied to the context of its production; it should therefore be understood as "an expression of the coloniality of power" (Ricaurte, 2019, p. 351). Within this regime, data relations can be defined as "new types of human relations that enable the extraction of data for commodification" (Couldry & Mejias, 2019, p. 337). This extraction is achieved in such a way as to over-surveil particular marginalised citizens, expel particular people from the social order and to suppress or eradicate alternative viewpoints and epistemologies (Escobar, 2017; Santos, 2009). Drawing together these and other effects, Ricaurte (2019, p. 351) notes how this trend "has led to new forms of colonization through data, grounded in material infrastructures and symbolic constructions that reinforce these practices." Thus, the task of reorganising the power structures that underpin big data systems—such as data extraction, storage, processing and analysis—requires a much broader and more rigorous process that must be produced, from the outset, through a decolonial lens (Arora, 2019). Another proposal for tackling some of the more pressing ethical concerns around big data is to reverse the dominant narrative by sharpening the focus on 'small data.' As Lupton (2014, p. 4) has pointed out: "while critical data studies often focus[es] on big data, there is also need for critical approaches to 'small' or personal data, the type of information that people collect on themselves." This was the known intention, for example, of a 2015 special issue of GeoJournal, appropriately titled *What's So Big about Big Data?*. The collection took the "end of theory" as its starting point of provocation for analysing the "epistemic limits of Big Data and accentuating the emerging social, political, and analytic challenges posed by Big Data research and analysis" (Burns & Thatcher, 2014, p. 446).

CONCLUSION

Despite the intensifying critical attention that big data's privacy politics are attracting, it is important to remind ourselves that the contours of big data have changed in the past and will continue to change, particularly with regards to shifting definitions of what privacy means both at an individual and a societal level. Ideas around what privacy means in relation to data must be continually reconsidered via the lens of questions of access, equity, ethics and accountability. Most importantly, this process needs to foreground the assumed 'bigness' of big data by challenging the notion that 'bigger' is necessarily better when it comes to research and knowledge production. A useful way out of this dilemma is to provide ethical and meaningful frameworks that navigate the complex connections between data and social phenomena. Embracing this will not only work to foreground the now generally-accepted notion that data is not ever merely 'raw' material, as well as assisting to build more ethical data infrastructures and analytical methods into our day-to-day practices as researchers, teachers, practitioners and consumers.

8. The Data Justice Lab does, however, acknowledge that in some cases the data harm examples listed could fit into several categories simultaneously (Redden & Brand, 2018).

REFERENCES:

Andrejevic, M. (2014). Big Data, Big Questions: Big Data Divide. *International Journal of Communication*, (8), 1673-1689.

Arora, P. (2019). Decolonizing Privacy Studies. *Television & New Media*, 20(4), 366-378. doi.org/10.1177/1527476418806092

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. London, UK: The MIT Press.

Browne, S. (2015). *Dark Matters: On the Surveillance of Blackness*. North Carolina, USA: Duke University Press.

Burns, R., & Thatcher, J. (2014). Guest Editorial: What's so big about Big Data?: Finding the spaces and perils of Big Data. *GeoJournal*, 80(4), 445-448.

Cheney-Lippold, J. (2017). *We Are Data: Algorithms and the Making of Our Digital Selves.* New York: New York University Press.

Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, 20(4), 336-349. doi.org/10.1177/1527476418796632

Dencik, L., Hintz, A., Redden, J. and Warne, H. (2018) Data Scores as Governance: Investigating uses of citizen scoring in public services. Retrieved from https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf

Doreski, W. (1999). *Robert Lowell's Shifting Colors: The Poetics of the Public and the Personal*. Athens, USA: Ohio University Press.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. London, UK: St Martin's Press.

Escobar, A. (2017). *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. North Carolina, USA: Duke University Press.

Ferguson. AG. (2015). Big Data and Predictive Reasonable Suspicion. *University of Pennsylvania Law Review*, 163(2): 327-410.

Ferguson. A. G. (2017). *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York, USA: NYU Press.

Foucault, M. (1975). *Discipline and Punish: The Birth of the Prison*. New York, USA: Pantheon Books.

Gangadharan, S. P. (2012). Digital inclusion and data profiling. *First Monday*, 17(5).

Gellert, R., & Gutwirth, S. (2013). The legal construction of privacy and data protection. *Computer Law and Security Review*, 29(5): 522-530.

Goffman, E. (1969). Strategic Interaction. Philadelphia, USA: University of Pennsylvania Press.

González Fuster, G. (2014). *The Emergence of Personal Data Protection as a Fundamental Right of the EU*. Heidelberg, Germany: Springer.

Harford, T. (2014). Big data: A big mistake? *Significance: Royal Statistical Society*, 11(5): 14-19. doi.org/10.1111/j.1740-9713.2014.00778.x

Heeks, R., & Renken, J. (2018). Data justice for development: What would it mean?. *Information Development*, 34(1), 90-102.

Hintz, A., Dencik, L., & Wahl-Jorgensen, K. (2017) Digital Citizenship and Surveillance — Introduction. *International Journal of Communication*, 11, 731-739.

Ioannidis, J. P. A. (2013). Informed Consent, Big Data, and the Oxymoron of Research That is Not Research. *The American Journal of Bioethics*, 13(4), 40-42.

Johnson, J. A. (2014). From Open Data to Information Justice. *Ethics and Information Technology*, 16(4), 263-274.

Koops, B-J. (2014). The trouble with European data protection law. *International Data Privacy Law,* 4(4), 250-261.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24), 8788-8790.

Kulhari, S. (2018). *Building-Blocks of a Data Protection Revolution: The Uneasy Case for Blockchain Technology to Secure Privacy and Identity.* Baden-Baden, Germany: Nomos.

Lupton, D. (2014). You Are Your Data: Self-Tracking Practices and Concepts of Data. Retrieved from http://papers.ssrn.com/sol3/Papers. cfm?abstract_id=2534211.

Lyon, D. (2003). Surveillance as Social Sorting: Computer Codes and Mobile Bodies. In D. Lyon (Ed.), *Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination* (pp. 13-30). London, UK and New York, USA: Routledge.

Mann, M., & Daly, A. (2018). (Big) Data and the North-in-South: Australia's Informational Imperialism and Digital Colonialism. *Television & New Media*, 20(4), 379-395. doi.org/10.1177/1527476418806091

Matzner, T. (2014). Why privacy is not enough privacy in the context of "ubiquitous computing" and "big data". *Journal of Information, Communication and Ethics in Society*, 12(2), 93-106.

Mayer-Schönberger, V., & Kenneth, C. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York, USA: Houghton Mifflin Harcourt.

Mosco, V. (2015). *To the Cloud: Big Data in a Turbulent World*. Boulder, UK: Paradigm.

Mosco, V. (2016). Marx in the Cloud. In V. Mosco (Ed.), *Marx in the Age of Digital Capitalism*. Leiden, Nederland: Brill.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, USA: NYU Press.

Parsons, C. (2015). Beyond Privacy: Articulating the Broader Harms of Pervasive Mass Surveillance. *Media and Communication*, 3(3), 1-11. doi.org/10.17645/mac.v3i3.263

Powles, J., & Nissenbaum, H. (2018). The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. Retrieved from https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

PrivazyPlan. (2018). Article 6 EU GDPR "Lawfulness of processing." Retrieved from http://www.privacy-regulation.eu/en/article-6-lawfulness-of-processing-GDPR.htm

Redden, J. (2018). The Harm That Data Do. Retrieved from https://www.scientificamerican.com/article/the-harm-that-data-do/

Redden, J., & Brand, J. (2018). Data Harm Record. Retrieved from https://datajusticelab.org/data-harm-record/

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. *Television & New Media*, 20(4), 350-365. doi.org/10.1177/1527476419831640

Richterich, A. (2018). *The Big Data Agenda: Data Ethics and Critical Data Studies.* London, UK: University of Westminster Press.

Rothstein, M., & Shoben, A. (2013). Does Consent Bias Research? T*he American Journal of Bioethics*, 13(4), 27-37.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2016). Automation, Algorithms, and Politics | When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software. *International Journal of Communicatio*n, 10, 4972-4990.

Santos, B. (2009). *An epistemology of the South: the reinvention of knowledge and social emancipation*. Mexico City, Mexico: Siglo XXI.

Schroeder, R. (2018). Social Theory after the Internet: Media, Technology, and Globalization. London, UK: UCL Press.

Schroepfer, M. (2018). An Update on Our Plans to Restrict Data Access on Facebook [Web log post]. Retrieved from https://newsroom.fb.com/news/2018/04/restricting-data-access/

Taylor, L. (2017). What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2918779

Tene, O., & Polonetsky, J. (2012). Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property*, 239, 243-251.

van der Ploeg, I. (2007). Genetics, biometrics and the informatization of the body. *Annali dell'Istituto Superiore di Sanita*, 43(1), 44-50.

Warren, S., & Brandeis, L. (1890). The Right to Privacy. *Harvard Law Review*, 193(4), 193-220.

Westin, A. F. (1967). Privacy and Freedom. *Washington and Lee Law Review*, 25(1), 7.

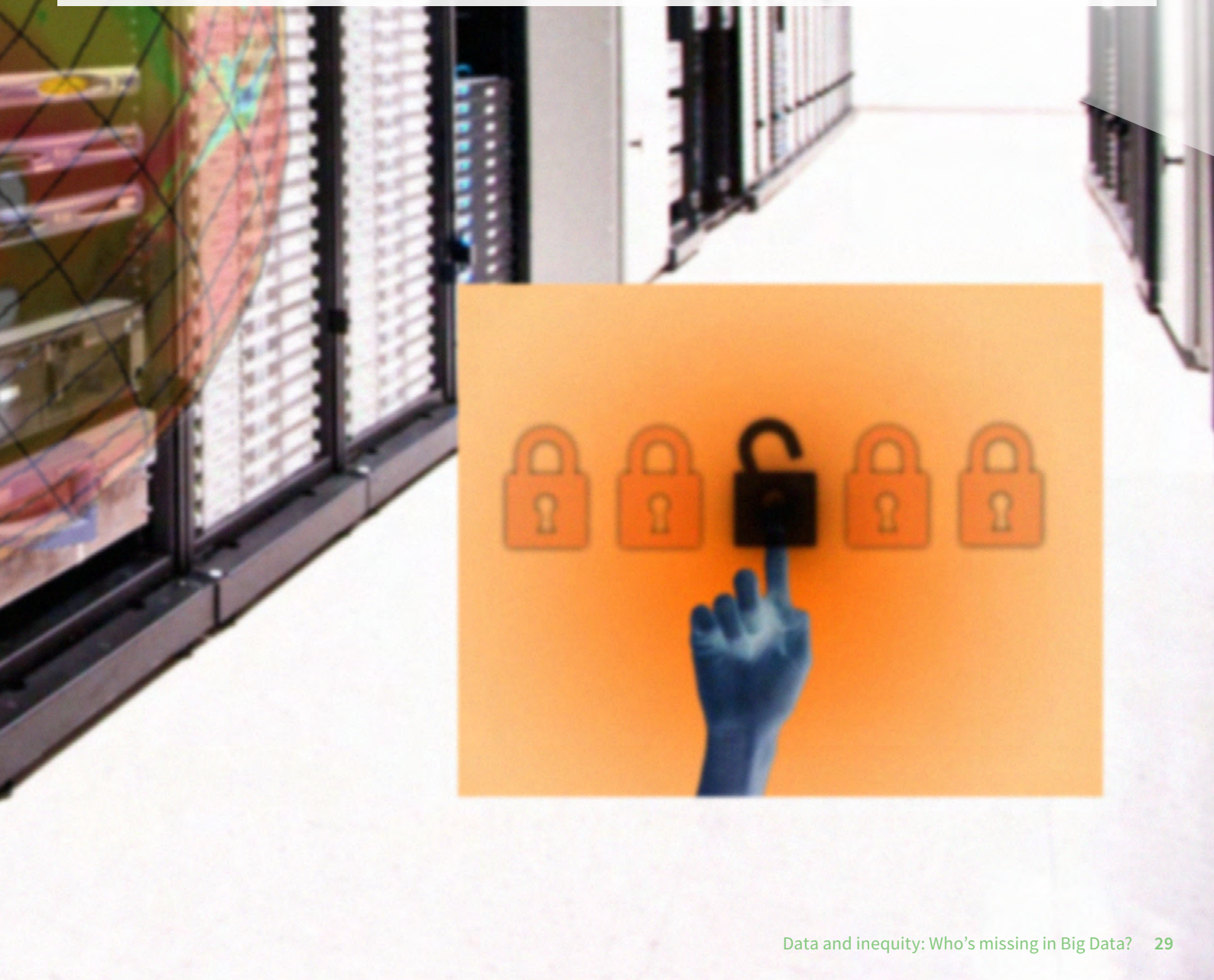
DATA AND INDIGENOUS PEOPLE

By Amba-Rose Atkinson, proud Gumbaynggirr, Wakka Wakka and Kabbi Kabbi woman from NSW; living, working and studying on Wurundjeri country.

ABSTRACT

What is data and what are its relationships to Aboriginal and Torres Strait Islander people in Australia?

Data can be defined as information that record the literal occurrences of our external world. The way in which data have been collected, presented, stored and used with regard to Indigenous Australia— a collective term used in the context of this chapter to refer to both Aboriginal and Torres Strait Islander people—has denied us ownership, control and distribution of this data, whilst inextricably entrenching our narrative in a discourse of deficit. The desensitised misrepresentation that data are merely numbers and not reflective of the lived experiences of Indigenous Australians has helped to maintain, throughout Australia's colonial history, a rhetoric of dispossession and domination. Resistance, however, has persevered and an ever-growing field of Indigenous academics and thinkers has created a new wave of data sovereignty empowerment. Ultimately this means we now have more control over how we give data, who can collect it, how it is disseminated and what it is used for. There are still, however, ongoing implications regarding data usage and inequity regarding Indigenous Australians. To contextualise this from an Indigenous standpoint, this chapter will briefly discuss the historical misuse of data and how it has positioned Indigenous people in Australia. The remaining discussion will focus on Indigenous data sovereignty; a central theme to the Indigenous panel, which comprised of Professor Marcia Langton, Dr Lyndon Ormond-Parker, Darren Clinch and Dr Tess Ryan as acting Chair.



BRIEF HISTORICAL OVERVIEW

Arguably, instances of historical misuse and non-consensual collection of data representing Indigenous people in Australia can be directly linked to the systematic disadvantage and social exclusion experienced by Indigenous Australians today. Currently, the relationship between Indigenous Australians and data and research remains tentative in some research spaces, highlighting the importance of creating platforms for Indigenous narratives and ownership of data. In this next section I will briefly examine key systems, theories, and policies that failed to uphold both ethical research practices and the basic human rights of Indigenous peoples in Australia.

Example of Historical Misuses and Collection of Data

Being the unwarranted and non-consensual object of scientific research has caused a great deal of pain in Indigenous communities around Australia, particularly from archival and research institutions such as museums, universities and art galleries, both within Australia and internationally. For example, in Australia, it is estimated that there are approximately 10,000 Aboriginal remains being kept within museums; internationally in places such as the United Kingdom, Germany, France and the United States of America, experts have estimated that there are approximately 1,000 Aboriginal remains being withheld (Korff, 2019). The underlying premise of obtaining and withholding Aboriginal remains has historically been based on the insistence that Aboriginal bones contained "unique evidence," and that it was the "duty" and "role of the museum" to "further the pursuit and dissemination of knowledge" (Jenkins, 2011, p. 1-3), despite the resistance and outcry from Indigenous communities around Australia, and the world. This example demonstrates the blatant disregard that non-Indigenous researchers have had for both the cultural and ceremonial significance placed on the remains of Indigenous people in Australia. It also demonstrates a blatant disregard for the ethics of conducting research on subjects who were neither given a choice, nor gave their consent for such use.

In recent times there are much more ethical guidelines as to how to conduct research, particularly if the research is focused on human beings as its subject. A chapter in *Researching Indigenous health: A practical guide for researchers*, published by The Lowitja Institute (2018), discusses how non-Indigenous researchers have exploited Aboriginal people for their knowledge about surrounding ecosystems, ranging from plants and wildlife to specific land uses. The chapter explains how non-Indigenous researchers would extract this rich and useful information and take it back to their universities for their own self-fulfilling agendas, with little regard for the Aboriginal people from whom their research had benefitted and often profited. In one section, the document states that an anthropologist was given "sensitive knowledge or objects by Elders helping with the research, then either published the information or displayed the sacred object in a public place", a practice which "risked the safety of the custodians who were seen to be breaking Aboriginal law" (The Lowitja Institute, 2018, p. 7). The inherent issue with this situation is a lack of cultural and intellectual property protection which, if instituted, could have held such non-Indigenous researchers accountable to the way in which they acquired, used and disseminated such material.

Critically acclaimed lawyer and legal firm owner Terri Janke—a proud Wuthathi/Meriam woman from Cairns (Janke, 2019)—describes intellectual property regarding traditional Indigenous knowledges as being the "rights Indigenous Australians have to their heritage [where] such rights are also known as Indigenous Heritage Rights" (Janke & Frankel, 1998, p. XVII). Published in the report *Our Culture: Our Future: Report on Australian Indigenous Cultural and Intellectual Property*, the term 'heritage' holistically encapsulates "the intangible and tangible aspects of the whole body of cultural practices, resources and knowledge systems developed, nurtured and refined by Indigenous people and passed on by them as part of expressing their cultural identity" (Janke et al., 1998, p. XVII). Notably, the term heritage is specified as: "languages; scientific, agricultural, technical and ecological knowledge (including cultigens, medicines, and sustainable use of flora and fauna); Indigenous ancestral remains and Indigenous human genetic material", amongst other items (Janke et al, 1998, p. XVII). Janke (1999, para. 15) notes that "Indigenous cultural and intellectual property rights are fundamental to the continuation and maintenance of Indigenous culture." According to Kukutai and Taylor (2016, p. xxi), data can be generated to measure and track the ways in which:

The rights of Indigenous peoples' access and ownership of lands, territories and resources are being met; how their participation in decision-making and control over their own development processes are progressing; what control over data and knowledge they are achieving; and what discrimination and exclusion they experience in regard to their social, economic and cultural rights.

What this ultimately means is that in today's research environment, data has the potential to be used as a positive tool to generate information that can be used to empower Indigenous communities around Australia.

Example of Key Systems

The notion of 'eugenics' has been used to justify genocide around the world, including in Australia less than two generations ago. The term eugenics was described by Galton and Galton (1998, p. 99), as "the science of improving inherited stock, not only by judicious matings, but by all the influences which give more suitable strains a better chance." The term is plausibly harmless when applied to the study of flora, however in the context of Indigenous Australians the term has been used to condone the forcible and calculated separation of infants and children from their families and communities, whereby most were placed in institutions to be taught how to be domestic servants for settler Australians. The intention was to implement a system that classified the blood quantum—the amount of 'full-blood'—of the Indigenous people of Australia, in an effort to "breed out the black" (Melville, 2018, para. 9). This system was favoured by settler Australians where they fervently believed they simply had to "await the "natural" death of the "full-blood" peoples and to socially engineer the disappearance, forever, of all those "natives of Aboriginal origin"" (Tatz, 1999, p.28). The language that was used to describe Indigenous people in Australia at this time was utterly void of humanity. By classifying people by the amount of 'blackness' in their DNA, what was being said was that to be black was to not be worthy of basic human rights; that the colour of one's skin—or lack of—could dictate how much dignity and respect one is entitled to. Further, such language reduces the human experience of being an Indigenous person to the status of a number and into a thereby quantifiable narrative. When today people are confused by the fact that it is hurtful to ask us, for example, "what percentage are you?" or "you don't look fully Indigenous, you must have something else?", what this essentially does is reinforce the colonial rhetoric that was used to denigrate and support genocidal policies that occurred less than two generations ago.

Example of Key Theories

Within the world of data and research, Indigenous Australians have been consistently excluded from the conversation; they have been denied contribution to their own narrative, which ultimately problematises the outcome of research and generates an unsafe research environment for Indigenous Australians. Moreover, it can be suggested that the traditional way of conducting research from a Western European perspective is deeply entwined within the undertones of epistemological racism (Bodkin-Andrews & Carlson, 2014, p.784). Bodkin-Andrews and Carlson (2014, p. 793) summarise epistemological racism as pertaining to research and its corresponding methodologies, theories and ways of knowing that have ultimately emerged from the social history of a dominant group, effectively diminishing and disregarding the perspectives of Indigenous peoples.

Example of Key Policies

The classification, removal of children and forced assimilation of the Indigenous people of Australia was enabled through government policies. Throughout the nineteenth and twentieth centuries, states and territories around Australia made it their mission to seek unsolicited and unwarranted control of the Indigenous people of Australia. For example, when Victoria devised the *Aboriginal Protection Act 1869*, it was implemented to control every aspect of the lives of Victorian Aboriginal people. This included constant pervasive observation and the punitive dictating of where Victorian Aboriginal people could live, what language they could speak, what they could eat, where they could work, and notably, who they could marry and have children with (Museum of Australian Democracy, 2019, para. 1). What ensued from this assimilative ideology was mass cultural dispossession, geographic displacement and violent institutionalisation. Today, we recognise those children who were forcibly removed from their families and communities across Australia as the Stolen Generation. As Beresford (2012, p. 65) notes, the overarching theme across all historical policies regarding Indigenous Australians was designed to "eradicate Aboriginal culture through assimilation" into mainstream Australian society. Today, when it comes to examining relationships between Indigenous Australians and the usage of data and research, is it important to recognise that we have been positioned to have to reclaim our narratives, knowledges and identities from within a space that has repeatedly attempted to control and erase us.

THE 2018 DATA SYMPOSIUM

In November 2018, the University of Melbourne and the Data, Systems and Society Research Network hosted a data symposium. The Indigenous Panel comprised of Professor Marcia Langton, Dr Lyndon Ormond-Parker and Darren Clinch with Tess Ryan as acting Chair. The focus of the panel was about the relationship and positionality of Indigenous Australians within the data and research space. The three panellists spoke about varying elements of research and the implications of these for Indigenous Australian people. While each panellist maintained a diverse dialogue, there were common themes that translated throughout each presentation. These can be noted as: the reliability of data surrounding Indigenous Australians; the need to improve data literacy of Indigenous Australians; the need to improve data protection and access; combating deficit discourse; and the importance of data in preserving and sustaining cultural knowledge. The first presentation will focus on Indigenous Data Sovereignty, the second on repatriation, preservation and protection; and the third presentation will focus on interconnecting modern technology and traditional art.

Indigenous Data Sovereignty

Firstly, the concept of Indigenous data sovereignty is described by Kukutai and Taylor (2016, p. xxii) as being "linked with Indigenous peoples' right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions, as well as their right to maintain, control, protect and develop their intellectual property over these." The first panellist to present and discuss Indigenous data sovereignty was renowned Aboriginal academic, Professor Marcia Langton AM. Langton is an anthropologist and geographer who has, since 2000, held the Foundation Chair of Australian Indigenous Studies at the University of Melbourne. In 1993, Langton was made a member of the Order of Australia in recognition of her work in anthropology and the advocacy of Aboriginal rights. In 2016 she was honoured as a University of Melbourne Redmond Barry Distinguished Professor and in 2017 was appointed as the first Associate Provost at the University of Melbourne. Langton is a descendant of the fighting Yiman people from central eastern Queensland.

Langton's (2018) presentation provided a summary and account of the Indigenous Data Network (IDN): an initiative of the University of Melbourne established in 2017 to realise the rights of Indigenous people to govern their own data in order to inform further developments, allocate resources and set future goals and objectives for themselves. In setting these objectives, the project seeks to build expertise and capacity to work more effectively and collaboratively with Indigenous communities and organisations across both Australia and internationally. Subsequently, IDN seeks to inform how future research can be undertaken within and outside of the academy. The project aims to provide clear opportunities for post-graduate students and early career researchers to engage in genuinely reciprocal and sustained partnerships with Indigenous communities, raising the profile and promoting a more enriched understanding of Indigenous culture, knowledge and values (Langton, 2018).

Langton (2018) continued by explaining that in 2018, the majority of the Australian Government's targets for the Close the Gap (CTG) campaign to improve Indigenous health, education and social participation had not been met—a clear indicator of the urgent need for innovative, paradigmatic shifts in how to address Indigenous disadvantage. Langton believes that a critical focus on Indigenous data sovereignty is one such way to change our national 'deficit thinking' approach, which would allow Indigenous people to set their own agendas, by using data about themselves to secure their social, economic, cultural and health-related futures.

One key point from Langton's (2018) presentation highlighted the need to bring "culture into the academy." University and similar research settings can ensure this by installing frameworks and protocols within faculty to ensure the increased presence of Indigenous researchers and professionals. When Indigenous researchers and professionals are at the helm of investigating the status of their own people, it reduces the risk of producing data that represents deficit discourse. Currently, the status of Indigenous Australians—particularly in the health space—is largely reported in a negative way, where Indigenous people are showcased as "failing" (Langton, 2018). As Langton made mention of, one such example of this is the annual CTG Government Report. Black, Pholi and Richards (2009), of the University of Newcastle, believe the campaign is another example of the type of wishful thinking around the power of statistics that seems to be at work in Indigenous policy circles in Australia. The thinking behind the campaign is that the presence of improved and increased data will somehow translate into improved and increased health and wellbeing for Aboriginal and Torres Strait Islander people. In a 2018 article published online by Probono Australia (Michael, 2018) the CEO of the Victorian Aboriginal Community Controlled Health Organisation (VACCHO), stated that "the strategy really needs to engage the Aboriginal community in it, and the Aboriginal community also has to lead it. And then governments should look at what it can do to achieve this greater outcome."

In order to ensure that data regarding Aboriginal and Torres Strait Islander people is being used appropriately—especially in nation-wide reports such as the CTG report—governments and research institutions alike need to prioritise employment and consultation with Indigenous organisations and researchers who are working in Indigenous affairs. Similarly, factually representing the overall status of Aboriginal and Torres Strait Islander people in such reports will make progress in better informing which areas require short-term versus long-term attention, with a more informed allocation of corresponding resources.

Langton's (2018) presentation was of paramount importance in demonstrating to non-Indigenous researchers why and how the narrative surrounding Aboriginal and Torres Strait Islander people must shift to include sovereignty, ownership and distribution of Indigenous data. At the end of the day, Indigenous researchers, professionals, Elders, and community members are the ones who know how to best ensure their data is used and distributed appropriately.

Repatriation, Preservation and Protection

The second panellist to present as part of the Indigenous Panel was Dr Lyndon Ormond-Parker, an ARC Research Fellow in the Indigenous Studies Unit at the Centre for Health Equity in the Melbourne School Population and Global Health, University of Melbourne. His ARC-funded research is currently focused on the Aboriginal Remote Narrowcast TV and the Audiovisual Archive (2018-2021). Ormond-Parker's areas of expertise relate to Indigenous cultural heritage, information technology and Indigenous communities. He is a member of the Australian Heritage Council as well as being a member of the Advisory Committee on Indigenous Repatriation. Ormond-Parker was born in Darwin and is of Alyawarr decent from the Barkly tablelands region of the Northern Territory.

Ormond-Parker's (2018) presentation, *Aboriginal Community Archives as Big Data*, explored issues pertaining to remote community archives as big data, looking at how factors of remoteness and inequality impact on the preservation of community languages, history and culture. His presentation looked at how information technology decisions are being made regarding hardware, networks and software; at the use of data management for specific purposes such as digital archives and access; and at how data governance as a community responsibility impacts on the longevity of digital data in remote Aboriginal communities.

Aboriginal and Torres Strait Islanders at the community and individual levels become more culturally empowered when their data is repatriated. It is a strange paradoxical phenomenon whereby Indigenous Australians must rely on the vast number of recordings taken by the settler society—that is, using a Western construction of Indigenous language, culture and identity—to therefore learn about our own culture. A large portion of Indigenous communities around Australia have been denied access to, and practice of, their own cultural and intellectual property; repatriating, reclaiming and revitalising Indigenous knowledge is therefore a necessary step in the survival and continuation of Indigenous cultures.

In Wadeye—a remote community in the Northern Territory that Ormond-Parker has worked in—there are many tapes on the history of the community, including the ceremonies that took place there and the oral languages spoken. It is of uttermost importance that the Aboriginal people in Wadeye, and other communities around the country, have their cultural and intellectual property control returned to them, by storing their data in sustainable data archives that can sensitively respond to the evolving research environment (Byrne, 2009, p. 1). In doing so, the people of Wadeye would be able to keep their knowledge safe, making it more accessible to share within their community and to have available for future generations. Ormond-Parker (2018) went on to propose a series of challenges in doing so, particularly around how communities might effectively retain control over their data and the logistics of that data's protection. He believes that by involving key stakeholders such as the Aboriginal community, dedicated researchers and national research institutions, there can be positive effects on the longevity and maintenance of data storage in remote Aboriginal communities (Ormond-Parker, 2018).

This presentation was a significant contribution to the discussion on how data archives have the opportunity to play an empowering role in the repatriation, preservation and protection of Aboriginal and Torres Strait Islander people and their cultures. As evidenced in the work that Ormond-parker was involved in throughout remote parts of Western Australia and the Northern Territory, it is imperative that Indigenous people at the community and individual level have access to interpreting, storing and sharing their own data. Such practices have the ability to shift responsibility and ownership back into Indigenous people's communities, in order for them to continue being keepers of their knowledge; this is precisely what empowers communities towards sustaining and strengthening their cultural traditions.

Interconnecting Modern Technology and Traditional Art
The final speaker to present as part of the Indigenous Panel was Darren Clinch, a proud Badimia man from Yamatji country in the mid-west of Western Australia. Clinch's current role involves developing business intelligence solutions for program areas within which the Department of Health and Human Services aim to leverage their powers regarding the departments' vast quantities of data and information for both reporting and monitoring, analytics and narrative style dashboards. Included in this work is the development of an Aboriginal Information System that can enable users to interpret patterns, trends and relationships within datasets that are presented using 'associative data modelling' visualisations. Overall, one of Clinch's key aims is to promote Indigenous Data Sovereignty and Governance principles and practices.

Clinch believes that utilising Aboriginal art and symbols to navigate data and technology increases its capacity to reach an Indigenous audience, particularly in the health space in relation to promoting and sending out health messages. For Clinch (2018), it is about "getting data about Aboriginal people into the hands of Aboriginal people and Aboriginal organisations." Too often have non-Indigenous researchers nullified the lived experiences of Aboriginal people, whereby the data that is used is not reflective of a holistic social situation. Similar to Langton's expression of dismay at the lack of transparency and accuracy in reports such as CTG, Clinch also questioned the reliability of data usage regarding Indigenous people in Australia.

For Clinch (2018), Indigenous data sovereignty resembles a space in which Indigenous researchers and professionals are behind innovative data practices, such as the one he himself is working on. As part of this work Clinch is combing Aboriginal art and technology practices to address data literacy and accessibility through the use of language and visuals that are familiar to Indigenous audiences. As he has noted, "my passion is not data for data's sake, but the value that data can bring to a story" (Clinch, 2018).

CONCLUSION

In concluding, we can start to form a picture of how essential Indigenous data sovereignty is in the lives of Indigenous Australians. All three panellists unanimously expressed the ways in which Indigenous data sovereignty is critical to a more successful inclusion of Aboriginal and Torres Strait Islander people, particularly within spaces in which we have long been misrepresented and silenced.

The Indigenous Panel as part of the 2018 Data Symposium was a testimony to the growing field of Indigenous academics and thinkers who are championing data sovereignty empowerment. It is evident that if we continue to include Aboriginal and Torres Strait Islander people in the discussion—particularly those conversations regarding Aboriginal and Torres Strait Islander people directly—then we find that future directions of the data and research space can finally become a more equitable and inclusive field of practice. Individually and collectively our voices are essential to this conversation; it is time that we are heard.

REFERENCES:

Beresford, Q. (2012). Separate and Equal: And Outline of Aboriginal Education. In Q. Beresford, G. Partington & G. Gower (Eds.), *Reform and Resistance in Aboriginal Education* (pp. 1-498). Western Australia: University of Western Australia.

Black, D., Pholi, K., & Richards, C. (2009). Is 'Close the Gap' a useful approach to improving the health and wellbeing of Indigenous Australians? *Australian Review of Public Affairs*, 9(2), 1-13.

Bodkin-Andrews, G., & Carlson, B. (2014). The legacy of racism and Indigenous Australian identity within education. *Race Ethnicity and Education*, 19(4), 784-807. doi.org/10.1080/13613324.2014.969224.

Byrne, A. (2009). The importance of culture in digital ecosystems: managing Indigenous data research. *MEDES 09' Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, 1-7.

Clinch, D. (2018). Data Sovereignty – How does it differ between government and the Aboriginal community controlled sector? *Data and Inequity: Who's Missing in Big Data? Data Symposium*. Retrieved from [insert link to recording].

Galton, C. J., & Galton, D. J. (1998). Francis Galton: and eugenics today. *Journal of Medical Ethics*, 24, 99.

Janke, T, & Frankel, M. (1998). Our Culture: Our Future: Report on Australian Indigenous Cultural and Intellectual Property Rights [Company Report]. Retrieved from https://d3n8a8pro7vhmx.cloudfront.net/ubcic/pages/1448/attachments/original/1485906181/Our_culture_our_future_report_2__copy.pdf?1485906181

Janke, T. (1999). Respecting Indigenous cultural and intellectual property rights. *University of New South Wales Law Journal,* 22(2), 631-639. Retrieved from http://www.austlii.edu.au/cgi-bin/sinodisp/au/journals/UNSWLJ/1999/16.html#Heading63

Janke, T. (2019). Terri Janke and Company: Lawyers & Consultants. Retrieved from http://www.terrijanke.com.au/team/terri

Jenkins, T. (2011). *Contesting Human Remains in Museum Collections: The Crisis of Cultural Authority* (First Edition). New York, USA: Routledge.

Korff, J. (2019). Aboriginal remains repatriation. Retrieved from https://www.creativespirits.info/aboriginalculture/people/aboriginal-remains-repatriation

Kukutai, T., & Taylor, J. (2016). Indigenous Data Sovereignty: Toward an Agenda. Retrieved from http://press-files.anu.edu.au/downloads/press/n2140/pdf/book.pdf?referer=2140

Langton, M. (2018). A summary of the purpose and goals of the Indigenous Data Network. *Data and Inequity: Who's Missing in Big Data? Data Symposium.* Retrieved from [insert link to recording].

Melville, K. (2018). The brutal legacy of Sister Kate's, a children's home with a mission to 'breed out the black.' Retrieved from https://mobile.abc.net.au/news/2018-07-29/brutal-legacy-of-sister-kate-indigenous-childrens-home/10034434?pfmredir=sm&fbclid=IwAR3FlA825KvXRJBD0oz2KhA637BVAOIaWhJF2gO2BVNX7sbaSe4f9UvMXIA

Michael, L. (2018). A vision for a culturally confident Aboriginal community. Retrieved from https://probonoaustralia.com.au/news/2018/01/vision-culturally-confident-aboriginal-community/

Museum of Australian Democracy. (2019). Aboriginal Protection Act 1869 (Vic). Retrieved from https://www.foundingdocs.gov.au/item-sdid-22.html

Ormond-Parker, L. (2018). Aboriginal Community Archives as Big Data. *Data and Inequity: Who's Missing in Big Data? Data Symposium*. Retrieved from [insert link to recording].

Tatz, C. (1999). Genocide in Australia – an AIATSIS research discussion paper. Retrieved from https://aiatsis.gov.au/sites/default/files/products/discussion_paper/tatzc-dp08-genocide-in-australia.pdf

The Lowitja Institute. (2018). Part A: Indigenous health research in context, chapter 1: setting the scene for research. *The Lowitja Institute*, 1-22.


1992 (11)

URBAN DATA AND ITS ROLE IN CREATING AND ADDRESSING INEQUITY

By Ishita Chatterjee, Soheil Sabri and Gideon Aschwanden

ABSTRACT

At the present moment there are unprecedented amounts of urban data being created and stored. Data range from census data that collected periodically, to satellite imagery that are more frequently updated, to transportation stream data that is collected and processed in real-time. Current technological innovations have the potential to enable data-driven policy formulation and decision-making, while evidence-based policies allow drafting of reliable development plans and modelling of comprehensive future scenarios. Despite an overwhelming amount of data being generated and stored, however, information is not readily available everywhere at the same depth or quality. Due to this missing data, many urban challenges such as natural and man-made disasters, the persistence of slum conditions in informal settlements, socio-economic segregation and unequal access to urban services remain unresolved. Currently underpinning this inability for data and technological platforms to address urban challenges are those issues pertaining to fragmented data, lack of capability of integration of big data with traditional data structures, as well as questions around data reliability. In this essay we will discuss those issues identified by members of the Data and Cities panel: Gideon Aschwanden, Soheil Sabri, Michele Acuto, Ishita Chatterjee and the participants of the 2018 Data Symposium. This panel, which was comprised of experts across areas of urban analytics, urban policy, urban science and informal settlements, deliberated primarily on the limits and opportunities of data-driven urbanism.

Panel members sought to address primary questions such as who might be missing from data representation, questions pertaining to unequal access to data as well as some of the effective ways that data can be collected and shared with multiple stakeholders. Another important concern that was voiced was regarding the poverty of urban science, which was identified as a weak connection between the various urban disciplines that ultimately limits comparisons across types of urban data.

Current technological innovations have the potential to enable data-driven policy formulation and decision-making, while evidence-based policies allow drafting of reliable development plans and modelling of comprehensive future scenarios.

DATA-INFORMED URBANISM TO DATA-DRIVEN URBANISM

For a long time urban data has been generated to study cities and their citizens, in order to analyse the processes and connections that are formed between them. Such findings have been collected and analysed by governments, institutions and businesses alike to attempt to understand how cities work. This is done in order that plans can be made for future urban scenarios and in finding solutions to potential challenges such as natural and man-made disasters, the persistence of slum conditions in informal settlements, socio-economic segregation and unequal access to urban services. Various forms of data collection on cities and their inhabitants are used to inform policy decisions and formulations, such as the drafting of development plans and the modelling of future urban scenarios. Such processes of using data inputs to analyse cities—a process which in turn aids the governance of a given city—are known as instances of data-informed urbanism (Kitchin, 2015, p. 2). It should be noted, however, that such processes prove to be expensive and time-consuming ways in which to collect and comprehend such data. In the majority of cases where such practices occur, data itself is static in nature, often showcasing only "snapshots of urban phenomen[a]" (Kitchin, Lauriault, & McArdle, 2017, p. 1). Hence while data has been actively used in formulating policies and drafting urban development plans, in earlier instances time lags between data collection and policy implementation proved to limit the efficiency of the method. Similarly, the gaps in data collection—where data was not widely available in a uniform manner—inhibited a comprehensive understanding of the issues being studied.

More recently, however, there have been changes in the processing speeds by which data can be collected and interpreted. This has led to a shift in the way that data can now more effectively influence the planning and decision-making of cities (Henke et al., 2016). The key differences between datasets produced in earlier times versus those generated more recently is with regards to their respective scales and rates of processing (Spielman, 2017). In this age of big data we are constantly surrounded by sensors and cloud-based storage systems; there is an unprecedented amount of data being created and stored—from census data that is collected periodically, to satellite imagery that is more frequently updated, to transportation data streams that are collected and processed in real-time. This scenario provides the basis for more efficient data interpretation, as well as greater insight into urban life at a much more granular level, with faster processing speeds across the entire longitudinal section of a city (Kitchin, 2016, para. 16). Consequently, urban governance and their services—based on a greater reliance on data systems to frame specific urban agendas—have begun to manage urban life through highly networked data systems. An example of this scenario is the public transport system in San Francisco, Routsey San Francisco, that uses sensor technology to optimise transportation services based on analysis of real-time information (Lee, Hancock, & Hu, 2014, p. 89). There is a worldwide shift occurring away from systems that are limited by data that suffers from considerable time lags between points of data processing and its implementation. Cities are moving towards systems that are capable of producing larger volumes of data able to be processed in real-time. Kitchin (2017, p. 46) calls this a transformation from data-informed urbanism to data-driven urbanism.

Big data has changed our views and understandings of different urban phenomena considerably. Big data is characterised by a huge amount of data (known as 'volume') with the possibility of being in real or near real-time (known as 'velocity'); it comprises different structures within a dataset (known as 'variety') with various accountability measures (known as 'veracity') and variable rate of data flow (known as 'variability'). Subsequently, these aspects are of no use (known as 'value') unless it can be turned into information (Gandomi & Haider, 2015, pp. 138-139). In the context of urban research, big data has been used to study different urban phenomena including housing affordability (Pettit, Tice, & Randolph, 2017), neighbourhood demography (Spielman, 2017) and accessibility to infrastructure (Benenson, Ben-Elia, Rofe, & Rosental, 2017). These capabilities have triggered a 'smart cities' movement in many countries globally (Thakuriah, Tilahun, & Zellner, 2017). The transition from data-informed to data-driven urbanism allows responsive urban governance and planning based on evidence (Kitchin, 2016, para. 15). Various scholars, however, have expressed doubts around such trends. Sennett (2012) is one such scholar sceptical about the prescriptive nature of data-driven smart city initiatives, whereas Vanky (2015) draws our attention to time lags between data input and their design outcomes. Examples of this include time lag between the information used for analysis of future scenarios and the urban interventions within a city based on that analysis. Kitchin and Lauriault (2014) on the other hand, have raised concerns about data privacy, control of

data and data security. While the symposium's Data and Cities panel deliberated on the limits and opportunities of data-driven urbanism, issues around privacy and security were taken up by a separate panel: one that was dedicated to Privacy.

While technological innovations have enabled data-driven policy formulation and decision making that can be supported by evidence, there are still challenges to big data that need to be addressed. Despite massive amounts of data being generated and stored, information is still not available everywhere at the same depth or quality. Due to this missing data, many urban questions pertaining to natural and man-made disasters, the persistence of slum conditions in informal settlements, socio-economic segregation and unequal access to urban services remain unanswered. Some of the issues underpinning the inability of current data platforms to address urban challenges relate to fragmented data, a lack of integration capability between big data and traditional data structures, as well as questions about data reliability. Since methods used for data analysis are constrained by the input of data, any biases present in the collected data are also transferred to the processing of data (Bondell, 2018) In recognising the implications of data usage as an incomplete representation of an entire population, the panel go on to discuss the role of data in creating inequity, drawing upon issues arising from both contexts of the Global North and South. Just as data practices have the ability to create inequity, they can also reveal inequity. The panel conclude their discussion by stressing the need for greater accountability of data practices to address such gaps present in current data inequity.

The first section discussed below speaks to issues of missing data, in particular the people and places that are missing from data-driven urbanism. The second section titled *Unequal Access to Data*, discussed the prevalence of digital gaps that restrict access to data, even when data is available. The third section on heterogenous and multi-sourced data considered problems related to data gathered by diverse sources, with methods suggested that might allow for better integration and collaboration possibilities between datasets. The final section, *Doing More with Data*, raised concerns over the poverty of urban science; that is, those weak connections between various urban disciplines that limit comparisons between urban datasets.

MISSING DATA

Despite massive amounts of data being generated and stored, the depth and quality of information is still disproportionately available in different parts of the world. There is a considerable imbalance in the availability of data when we compare, for example, cities between the Global North and South (Acuto, Parnell, & Seto, 2018, p. 3). The old saying "if you don't count it, it doesn't count" (in Cortright, 2016, para. 17) is of great relevance when considering issues of being left out of the datasets in this era of data-driven urbanism.

Around one billion people globally live in informal settlements, a critical mass that is for the most part invisible to their respective governments (U.N. Habitat, 2016, para. 1). Although there are noted benefits to this kind of invisibility—that are harnessed by both informal residents and the state—there are dramatic ramifications regarding the 'unmapped' (AlSayyad & Roy, 2004, p. 158) and 'non-notified' (Krishna, 2013, p. 1013) status of these settlements. A 'notified' status, on the other hand, refers to those settlements that are recognised by municipalities and local bodies within that area. While being notified does not necessarily provide inhabitants immunity from factors such as forced evictions, a non-notified status exacerbates social and health-related issues that exist within these settlements (Subbaraman et al., 2012). Without a fixed, known address, residents in these areas do not exist as part of official registers and are therefore unable to access basic services and subsidies intended for such populations living below the poverty line (Edelman & Mitra, 2006). Non-notified residents are also unable to procure official documents, making it very difficult for them to claim such basic human rights as a fresh water supply, electricity, sanitation, education and health facilities. Without such links to land and key services, policies that are aimed at improving the lives of the urban poor themselves fall short when this population group gets left behind. Such challenges are felt proportionately more in the wake of natural and other disasters, when this unaccounted for and most vulnerable group cannot be reached; in most cases these populations therefore miss out on various compensations promised by respective governments (Nolan, Bloom, & Subbaraman, 2017).

In addition to the deficiency of data, there are also problems regarding its homogenisation. Even in cases where informal settlements are reported—usually through either statistical or aesthetic calculative practices—these areas still go unaccounted for where both Global North and South epistemologies have 'slum-free' cities as part of their agendas. Globally speaking, there is a dominant approach to such poor areas which is to criminalise these settlements and their residents (Brugmann, 2009; Malecki & Ewers, 2007; Neuwirth, 2005). In his provocative essay *The Unintended City*, Sen (1975) points out that despite a city's dependence on such settlements, they are still perceived as a failed aspect of the urban fabric of a place. Being 'off the map' (Menon-Sen & Bhan, 2008) and sitting on a "zone absent of policies" until they are acknowledged (Subbaraman et al., 2012, p. 661) means that the on-ground realities of these settlements remain unrecorded. Within an Australian context, a similar debate was taken up during the Data and Indigenous People panel as part of the symposium, regarding the omission of an entire population of people from data practices. A common message that was shared by both the Data and Cities and Indigenous People panels was that decisions are made based on data that is collected, hence why partial data can give a very different picture of any situation, in many cases leading to erroneous results (Cortright, 2016).

In recognising the price of being left out in this data-driven age, inhabitants of informal settlements have been striving for visibility (Zimmer, 2012). Through community mapping and resident-driven data collection methods, settlement dwellers have started writing their own narratives against what are purposefully distorted population numbers being presented by government bodies and officials. Empirically grounded 'counter mapping' (Peluso, 1995)—stories like *Map Kibera* (Hagen, 2011), *Know Your City* (Byrne, 2018) and *Missing Maps* (Michael, 2014)—reveal that residents themselves are the underutilised resources needed for addressing these present gaps in datasets.

Along with the Global South, Global North cities are also faced with issues related to exclusion from data. For most countries, the amount of data being collected and produced in smaller cities is far less than what is being collected in bigger cities, proving there to be a definite metrocentric bias (Bunnell & Maringanti, 2010, pp. 416-417). As the panel pointed out, in Australia capital city municipalities are responsible for over 70% of published datasets.

While invisibility is one key reason for people being left out in this data-driven world, unequal access to data plays another major role in the creation of data gaps.

UNEQUAL ACCESS TO DATA

Central to issues of unevenly distributed access to data are questions of data ownership and control. In an age of overabundance of digitally-collected data, most of this information is being held by private companies including data collected by telecommunication operators, all forms of service providers and transport companies. Driven by a desire to compete in this data-driven world, such stakeholders either refuse to share their data or charge an exorbitant amount for access to this information. Hence the irony at play here is that even in an age of data revolution, access to a majority of data is vastly and unevenly limited (Kitchin, 2017, p. 51). Apart from factors of affordability and restrictions to access, other factors that contribute to this disproportionate access include issues related to data illiteracy. Distrust in technology as well as insufficient skills among users are some of the major determinants that contribute to the digital divide (Meijers, Stikker, & Schouten, 2018). This divide refers to uneven distribution in access to information and communication technology (ICT) as well as the inability to participate in civic life online and media due to digital illiteracy (Cohron, 2015). By extension, the digital divide restrains a group's participation in data-driven urbanism (Bott & Young, 2012; Tenney & Sieber, 2016).

Just as there are implications on the insights gained through the collection of data where gaps are prevalent, so too are there problems associated with abundant and incomparable datasets. The next section will delve into some of the technical issues related to data and urban data models.

HETEROGENEOUS AND MULTI-SOURCED DATA

With data being increasingly used in decision making processes, analysts and other decision makers are confronted with heterogeneous data that is being produced across incommensurable data sources (Rajabifard, Ho, & Sabri, 2016). For various organisations and institutions that explore ways of increasing accessibility of urban data, there are constraints evident in the disparities that exist between various unstructured types of data—for example crowd sourced data, as well as data contained in disciplinary silos, including environmental, planning and infrastructure groups (Rajabifard et al., 2016). Such organisations have their own methods for collecting, structuring, storing, analysing and distributing data according to various platforms used. In order to achieve successful outcomes, inputs are required from various disciplines and diverse data sources so that complex urban phenomena can be better understood. Data analysis across a range of platforms requires that such fragmented data be harmonised in order to then be used (Chen, Sabri, Rajabifard, & Agunbiade, 2018). As such, an efficient data infrastructure as part of this scenario would be one that enables data to be scalable, integrated and interoperable (Sabri, Rajabifard, Ho,

Namazi-Rad, & Pettit, 2015, p. 35). Spatial data infrastructures (SDIs) are a rapidly evolving concept designed to enable decision makers to successfully and accurately make decisions, by providing capabilities for spatial data to be accessed, integrated, processed and published (Rajabifard et al., 2016, pp. 97-100).

In Australia, the Australian Urban Research Infrastructure Network (AURIN) have been working at the forefront of this data-driven urban planning revolution by providing academics, governments and industry personnel access to a collaborative network at the national level. A recent research initiative by the Australian Research Council (ARC) is working with SDIs to augment AURIN and other data hubs' research capabilities; in Australia this has provided some answers to issues of integration regarding multi-source data. This initiative—known as the Urban Analytics Data Infrastructure (UADI)—is a collaborative effort between six Australian universities; collectively the aim is to develop an ontological framework that defines and relates concepts for the purposes of integrating multi-disciplinary datasets (Rajabifard et al., 2016, pp. 96-97).

In the Knowledge Engineering community "ontologies aim to capture consensual knowledge in a generic way, and that they may be reused and shared across software applications and by groups of people. They are usually built cooperatively by different groups of people in different locations" (Corcho, 2005, p. 4). Accordingly, ontological engineers formulate frameworks that concern developmental processes, life cycles, methodologies, tools and languages for building various ontologies.

DOING MORE WITH DATA

While the rethinking of ontological frameworks is necessary for the linking of different datasets, similarly, a cross-examination of the practices involved in data collection, analysis and its dissemination is also essential. Acts of collecting and organising data are never divorced from the presence and influence of ideologies, therefore data itself is never neutral information. Data "always shapes and is shaped" by the environment in which it is created and interpreted (Dalton & Thatcher, 2014, p. 3). Dalton and Thatcher (2017) call for urban data provenance by proposing an investigation into its source, context and history; by exploring "data about [data]," which we refer to as metadata (Kitchin et al., 2017, p. 5).

Adding to this conversation on metadata, the panel stressed the need for multi-stakeholder data, where diverse conclusions are able to be reached from the same datasets by different professionals. They pointed out that "data availability does not immediately translate into better-informed urban management," (Acuto, 2018, p. 165). They also questioned traditional practices of data usage that is geared towards end processes used in policy formulation, rather than such practices being the basis for more questions. There are of course benefits of "retrospective access to information," whereby research projects can discover new insights from looking at the datasets of an earlier research initiative (Sabri et al., 2015, p. 35). These discussions echoed what is known as FAIR data principles: Findability, Accessibility, Interoperability and Reusability; a model which ensures better data management and stewardship (Wilkinson et al., 2016).

At the heart of the exchange that took place between speakers as part of this panel was the realisation that much of the urban data that is available is insufficient in tackling what are some of the most pressing global challenges. While the bulk of the world's urbanisation is occurring in the Global South and within smaller cities generally, urban knowledge is still largely confined to cities of the Global North and metropolitan cities in general (McPhearson et al., 2016, p. 166). The threat of natural disasters and global warming is being felt at a global scale, while current urban scholarship is inadequate in understanding urban transformation at such a scale and pace (Acuto et al., 2018, p. 165-166). In order for there to be comprehensive understanding of the complex range of issues effecting urbanisation, cross-disciplinary approaches as well as stronger working relationships between academics and practitioners is required (Acuto et al., 2018, p. 3). In highlighting such weak interactions that exist between various urban disciplines—ones that limit effective comparison between sets of urban data—the panel spoke about the need for a better urban science.

One initiative they outlined in the move towards bridging of this gap was the establishment of the Nature Sustainability expert panel on science and the future of cities, which increases the capacity for scientific advice within the context of the United Nation's Sustainable Development Goals (SDGs) (Nature, 2018). The role of the international expert panel is to suggest productive ways in which to cultivate urban scholarship around policy engagement; this is achieved through addressing imbalances of capacity with respect to urban knowledge, issues related to access to data and the need for a critical approach to data drivers and their impacts (Nature, 2018). Armed with inputs from various disciplines, this urban science should be capable of giving sound policy advice, thereby revolutionising the ways in which urban knowledge is being created (Acuto et al., 2018, p. 4).

An example of an effective urban science is one that is able to foster productive, interdisciplinary collaboration; furthermore, it is one that is able to provide global perspectives on challenges of inequity, injustice and factors of socio-economic disparity (Parnell & Robinson, 2017, p. 21, 27). This new approach to urban science should be one that enables and encourages debate on data collection, usage and representation. A new data-driven urbanism should be a place for argument and critique; it should promote the robust use of argumentative science as well as evidence-based policy debate. By way of conclusion, the panel reminded us that data should not be the end point of a conversation, but rather the starting base for one.

REFERENCES:

Acuto, M. (2018). Global science for city policy. *Science*, 359(6372), 165-166.

Acuto, M., Parnell, S., & Seto, K. C. (2018). Building a global urban science. *Nature Sustainability*, 1(1), 2-4.

AlSayyad, N., & Roy, A. (2004). Prologue/Dialogue, Urban Informality: Crossing borders. In

AlSayyad, N., & Roy, A. (Eds.), *Urban Informality: Transnational Perspectives from the Middle East, Latin America, and South Asia*. Lanham, USA: Lexington Books.

Benenson, I., Ben-Elia, E., Rofe, Y., & Rosental, A. (2017). Estimation of urban transport accessibility at the spatial resolution of an individual traveler. In P. (V.)

Thakuriah, N. Tilahun & M. Zellner (Eds.), *Seeing Cities Through Big Data* (pp. 383-404). Switzerland: Springer.

Bondell, H. D. (2018). Are Biases Hiding in our Data and Algorithms? Lecture presented at the Data and Inequity: Who is Missing from Big Data? Symposium, The University of Melbourne, Australia. Retrieved from https://networkedsociety.unimelb.edu.au/__data/assets/pdf_file/0003/2929305/DSSRN-program_v5.pdf

Bott, M., & Young, G. (2012). The role of crowdsourcing for better governance in international development. *Praxis: The Fletcher Journal of Human Security*, 27(1), 47-70.

Brugmann, J. (2009). *Welcome to the Urban Revolution: How Cities Are Changing the World*. St Lucia, Australia: University of Queensland Press.

Bunnell, T., & Maringanti, A. (2010). Practising Urban and Regional Research beyond Metrocentricity. *International Journal of Urban and Regional Research*, 34(2), 415-420. doi.org/10.1111/j.1468-2427.2010.00988.x

Byrne, J. (2018). Know Your City: Slum Dwellers Count. Retrieved from http://knowyourcity.info/wp-content/uploads/2018/02/SDI_StateofSlums_LOW_FINAL.pdf

Chen, Y., Sabri, S., Rajabifard, A., & Agunbiade, M. E. (2018). An ontology-based spatial data harmonisation for urban analytics. *Computers, Environment and Urban Systems*, 72, 177-190. doi.org/10.1016/j.compenvurbsys.2018.06.009

Cohron, M. (2015). The Continuing Digital Divide in the United States. *The Serials Librarian*, 69(1), 77-86. doi.org/10.1080/0361526X.2015.1036195

Corcho, O. (2005). *A Layered Declarative Approach to Ontology Translation with Knowledge Preservation*. Amsterdam, The Netherlands: IOS Press.

Cortright, J. (2016). The limits of data-driven approaches to planning. Retrieved from http://cityobservatory.org/the-limits-of-data-driven-approaches-to-planning/

Dalton, C., & Thatcher, J. (2017). Data provenance and possibility: Thoughts towards a provenance schema for urban data. In R. Kitchin, T. P. Lauriault & G. McArdle (Eds.), *Data and the City* (Regions and Cities) (pp. 92-104). London, UK: Routledge.

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data.' Retrieved from https://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like-and-why-do-we-care-craig-dalton-and-jim-thatcher/

Edelman, B., & Mitra, A. (2006). Slum dwellers' access to basic amenities: The role of political contact, its determinants and adverse effects. *Review of Urban & Regional Development Studies*, 18(1), 25-40. doi.org/10.1111/j.1467-940X.2006.00109.x

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, 35(2), 137-144. doi.org/10.1016/j.ijinfomgt.2014.10.007

Hagen, E. (2011). Mapping change: Community Information Empowerment in Kibera (Innovations Case Narrative: Map Kibera). *Innovations: Technology, Governance, Globalization*, 6(1), 69-94.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). McKinsey Global Institute, The age of analytics: Competing in a data-driven world [Company Report]. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

Kitchin, R. (2015). Data-driven, networked urbanism. Retrieved from https://ssrn.com/abstract=2641802

Kitchin, R. (2017). Data-driven urbanism. In R. Kitchin, T. P. Lauriault & G. McArdle (Eds.), *Data and the City* (Regions and Cities) (pp. 44-56). London, UK: Routledge.

Kitchin, R. (2016). Urban big data. Retrieved from https://www.theplanner.co.uk/features/urban-big-data

Kitchin, R., & Lauriault, T. P. (2014). Towards critical data studies: Charting and unpacking data assemblages and their work. The Programmable City Working Paper 2. Retrieved from http://mural.maynoothuniversity.ie/5683/1/KitchinLauriault_CriticalDataStudies_ProgrammableCity_WorkingPaper2_SSRN-id2474112.pdf

Kitchin, R., Lauriault, T. P., & McArdle, G. (2017). Data and the city. In R. Kitchin, T. P. Lauriault & G. McArdle (Eds.), *Data and the City* (Regions and Cities) (pp. 44-56). London, UK: Routledge.

Krishna, A. (2013). Stuck in place: investigating social mobility in 14 Bangalore slums. *The Journal of Development Studies*, 49(7), 1010-1028.

Lee, J. H., Hancock, M. G., & Hu, M-C. (2014). Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco. *Technological Forecasting and Social Change*, 89, 80-99.

Malecki, E. J., & Ewers, M. C. (2007). Labor migration to world cities: with a research agenda for the Arab Gulf. *Progress in Human Geography*, 31(4), 467-484. doi:10.1177/0309132507079501

McPhearson, T., Parnell, S., Simon, D., Gaffney, O., Elmqvist, T., Bai, X., . . . Revi,

A. (2016). Scientists must have a say in the future of cities. Nature News, 538(7624), 165-166.

Meijers, E., Stikker, M., & Schouten, S. (2018). Lost in the numbers: The missing politics of Big Data. Retrieved fromhttps://www.greeneuropeanjournal.eu/lost-in-the-numbers-the-missing-politics-of-big-data/

Menon-Sen, K., & Bhan, G. (2008). *Swept off the map: surviving eviction andresettlement in Delhi*. New Delhi, India: Yoda Press.

Michael, C. (2014). Missing Maps: nothing less than a human genome project for cities. Retrieved from https://www.theguardian.com/cities/2014/oct/06/missing-maps-human-genome-project-unmapped-cities

Nature, S. (2018). Launching our first expert panel report. *Nature Sustainability*, 1(12), 721. doi:10.1038/s41893-018-0209-7

Neuwirth, R. (2005). *Shadow cities: A Billion Squatters, a New Urban World*. New York, USA: Routledge.

Nolan, L., Bloom, D. E., & Subbaraman, R. (2017). Legal Status and Deprivation in India's Urban Slums: An Analysis of Two Decades of National Sample Survey Data. *Econ Polit Wkly*, 53(15), 47-55.

Parnell, S. & Robinson, J. (2017). The global urban: difference and complexity in urban studiesand the science of cities. In S. Hall & R. Burdett (Eds.), *The SAGE Handbook of the 21st century city* (pp. 13-31). London, UK: SAGE Publications Ltd.

Peluso, N. L. (1995). Whose woods are these? Counter-mapping forest territories in Kalimantan, Indonesia. *Antipode*, 27(4), 383-406.

Pettit, C., Tice, A., & Randolph, B. (2017). Using an online spatial analytics workbench for understanding housing affordability in Sydney. In P. (V.)Thakuriah, N. Tilahun & M. Zellner (Eds.), *Seeing Cities Through BigData* (pp. 233-255). Switzerland: Springer.

Rajabifard, A., Ho, S., & Sabri, S. (2016). Urban Analytics Data Infrastructure: Critical SDI for Urban Management in Australia. In D. J. Coleman, A. Rajabifard, & J. Crompvoets (Eds.), *Spatial Enablement in a Smart World* (pp. 95-109). Gilbertville, USA: GSDI Association Press.

Sabri, S., Rajabifard, A., Ho, S., Namazi-Rad, M-R., & Pettit, C. (2015). Alternative Planning and Land Administration for Future Smart Cities [Leading Edge]. *IEEE Technology and Society Magazine*, 34(4), 33-73. doi.org/10.1109/MTS.2015.2494298

Sen, J. (1975). *The unintended city: An essay on the city of the poor.* Calcutta, India: Cathedral Relief and Social Services.

Sennett, R. (2012). No one likes a city that's too smart. Retrieved from https://www.theguardian.com/commentisfree/2012/dec/04/smart-city-rio-songdo-masdar

Spielman, S. E. (2017). The potential for big data to improve neighborhood-level census data. In P. (V.) Thakuriah, N. Tilahun & M. Zellner (Eds.), *Seeing Cities Through Big Data* (pp. 99-111). Switzerland: Springer.

Subbaraman, R., O'Brien, J., Shitole, T., Shitole, S., Sawant, K., Bloom, D. E., & Patil-Deshmukh, A. (2012). Off the map: the health and social implications of being a non-notified slum in India. *Environment and Urbanization*, 24(2), 643-663.

Tenney, M., & Sieber, R. (2016). Data-Driven Participation: Algorithms, Cities, Citizens, and Corporate Control. *Urban Planning*, 1(2), 101-113.

Thakuriah, P. V., Tilahun, N. Y., & Zellner, M. (2017). Introduction. In P. (V.) Thakuriah, N. Tilahun & M. Zellner (Eds.), *Seeing Cities Through Big Data* (pp. 1-9). Switzerland: Springer.

U.N. Habitat. (2016). Slum Almanac 2015–2016: Tracking Improvement in the Lives of Slum Dwellers. Retrieved from https://unhabitat.org/wp-content/uploads/2016/02-old/Slum%20Almanac%202015-2016_EN.pdf

Vanky, A. (2015). The Elusiveness of Data-driven Urbanism. In L. Sheppard & D. Ruy (Eds.), *103rd ACSA Annual Meeting Proceedings, The Expanding Periphery and Migrating Center* (pp. 177-185). Toronto, Canada: Association of Collegiate Schools of Architecture.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.

Zimmer, A. (2012). Enumerating the semi-visible. *Economic and Political Weekly*, 47(30), 89.

DATA AND HEALTH

Dr Kyle Turner, Professor Jodie McVernon and Dr Ruth De Souza

ABSTRACT

New forms of data and associated analytical methods are transforming healthcare, highlighting opportunities to achieve health equity for marginalised groups. Data obtained purposively in medical settings combined with new data sources from electronic medical records, other routine health data, social media, mobile applications and wearable technologies along with advances in analytical methods can better utilise the collective value of newly emerging information streams. These innovations include machine learning, multi-disciplinary partnerships and real-time analysis and forecasting (Stieb et al. 2017). However, there are barriers to realising equity given that 'big data' is often acquired through mobile devices, wearable technology or electronic medical records, that may be incomplete or exclude population groups with lighter digital footprints. For example, population subgroups may experience barriers to health because of their race, disability, sexuality, gender identity, socioeconomic status, access to care or health literacy. Thus big data can also reinforce injustice for those populations who already experience a disproportionate health burden, and may already be underserved by the health system. In the health panel, three case studies explored how to address the issue of missing data and the ways in which populations might benefit from addressing such a gap. In the first presentation Anne Kavanagh proposed mechanisms for ensuring the completeness of data for people with disabilities. The National Disability Insurance Scheme (NDIS), a personalised funding scheme for people with disabilities, is a timely example. Kavanagh proposed providing researchers with access to de-identified data, and providing capacity to link data with other resources in order to strengthen public health policy. This completeness could then also be used for advocacy for systems and service improvements. Karin Verspoor's presentation examined the potential benefits and pitfalls of Australia's new electronic My Health Record. This centralised, online repository used for the collection and access of health data, has the potential to improve the responsiveness of healthcare services for people, reducing the burden on consumers to be able to repeat their health histories. However, uptake has been controversial given concerns about privacy and the potential for data breaches, as well as the seemingly heavy handed automatic registration of consumers. While, Steven Tong's presentation demonstrated the power of data linkage for Aboriginal Australians by combining orphan datasets to reveal new insights and public health benefits that would have otherwise remained hidden from view. The question of how data are collected, managed and analysed remain a challenge. It is essential to ensure that data are representative and that equity considerations are addressed, and safeguards for sharing including data ownership, privacy and security are robust (Stieb et al. 2017).

INTRODUCTION

Big data has made many promises. In healthcare, it has been touted as showing the way forward for the next generation of drug discovery, treatment innovation and optimised patient care. Advocates for greater sharing of personal health data argue that the benefits to the individual and to the public far outweigh any potential concerns—notably patient privacy and security. The following three presentations showcase examples of what big data is capable of, while also highlighting the common apprehensions and roadblocks to accessing such increased volumes of patient data.

DEMOCRATISING DISABILITY DATA

Professor Anne Kavanagh is a Social Anthropologist who is best known for her work in the health inequalities arena. She is the current inaugural Chair of Disability and Health at the University of Melbourne's School of Population and Global Health where she is also Head of the Disability and Health Unit. She is also the Academic Director of the Melbourne Disability Institute and the Director and Lead Investigator on the Centre of Research Excellence in Disability and Health. For most of her career, Kavanagh's research focus has been on the health of people with disability. Her most recent focus has been on how social determinants such as employment, housing, poverty and education influence the health of people living with a disability. Her aspiration in doing this work is to identify policy solutions to assist in the reduction of disability-related socio-economic and health disadvantages in Australia and internationally.

Australia's largest minority group consists of those people living with a disability—both physical and intellectual—who make up approximately 18% of the total population. Despite this high number, these individuals are unfortunately not afforded the same basic human rights that other citizens take for granted (PwC, 2011). People living with a disability are more likely to experience poverty, live in poor quality and/or insecure housing and often have low-level educational attainment. These individuals are often socially isolated and have fewer available opportunities for taking part in community events and activities (Australian Human Rights Commission, 2019, para. 7). In the last five years, however, there has been significant demand for changes to the way that people living with a disability are treated. Most significantly, a bill to establish the National Disability Insurance Scheme (NDIS) was introduced into Federal Parliament in 2012 by then Prime Minister Julia Gillard (NDIS, n.d.). The bill was passed in 2013 under the title of the National Disability Insurance Scheme Act 2013, but a change in Federal Government soon after disrupted the scheme's implementation. According to Kavanagh, we have been playing catch up ever since.

The NDIS signals the largest Federal-level social policy reform since the introduction of Medicare in 1975. The Federal Government currently spends approximately $22 billion annually on the NDIS; this figure is predicted to balloon out to $30 billion by 2030 in order to provide services and support to 460,000 Australians with severe or permanent disability (2018, p. 26). In Kavanagh's view, however, the NDIS does not provide the appropriate necessary care for these one-in-five people, and as such, the Government is neglecting its duty as a signatory to the United Nations Convention on Rights of Persons with Disabilities (United Nations, 2006).

The Melbourne Disability Institute's flagship project—which is led by Kavanagh—is known as the Democratising Disability Data Coalition (DDDC). The primary function of the DDDC is to call out omissions of evidence in relation to the NDIS; its goal is to achieve safe and secure access to disability data and statistics in order to provide the evidence needed to optimise services and policy (Melbourne Disability Institute, n.d., para. 1). The DDDC's list of calls to action include:

- Making NDIS' de-identified data available to Data Integrating Authorities;
- Making governments accountable to their commitments to the ongoing Survey of Disability Ageing and Carers (SDAC);
- Making governments accountable to their commitments to the continued collection of the Report on Government Services (ROGS) Disability Data;
- Making sure that data belonging to the National Quality and Safeguards Commission is available and linked with the NDIS and other key databases (under the "Five Safes" protocol);
- Making sure that all disability services collect data on functional impairment of people with a disability, so that inclusion and access by people with a disability can be measured and monitored for the first time;
- Advocating for the next national Census to include a question to identify whether someone is an NDIS participant or not and;
- Making sure that other data in relation to disability marketplace and services is directly funded by governments and made available for research (Melbourne Disability Institute, n.d.).

A key concern regarding the release of NDIS data to research organisations such as the DDDC is the risk of re-identification and data breaching. There is, however, a counter argument in that these risks are far outweighed by the public health benefits to one of Australia's highest at-risk populations. This enduring debate between the potential benefits and risks posed by the collection of large datasets and how these practices can afford the healthcare sector continued into the next session presented by Professor Karin Verspoor.

RECOGNISING THE VALUE OF SHARING DATA

Professor Karin Verspoor is from the School of Computing and Information Systems at the University of Melbourne and is the Deputy Director of the Health and Biomedical Informatics Centre. Her research has focused primarily on extracting information from clinical texts and biomedical literature that uses machine learning (ML) methods. The focus of this presentation was on a topic that has been attracting headline visibility recently in Australia: My Health Record (Australian Digital Health Agency, n.d.).

My Health Record is a centralised, online repository used for the collection and access of health data. Its purpose is to provide various healthcare professionals with around the clock, ongoing access to a patient's personal health information from anywhere in the country. The primary aim of My Health Record is to allow patients and doctors access to timely medical information such as test results, referral letters and organ donation information (Margo, 2018). A secondarily function of the record is to allow academics access to aggregated and de-identified datasets for public health research (The Age, 2019). Similar to the challenges facing the NDIS, a number of concerns have been flagged around the safety of some of the more personal and sensitive patient data collected using My Health Record (Zhou, 2018).

One example of this concern stems from the Australian Government Department of Health having stated—prior to the launch of My Health Record—that any secondary use of patient data must be of public benefit and cannot be solely commercial, while private health insurance companies were not allowed to participate at all (Bogle, 2019, para. 19). This fact is not entirely problematic, except when you consider that the exact definition of this exclusionary clause for commercial interest remained at best vague. The framework body of the Department of Health that was overseeing the secondary use of My Health Record data was being reviewed at the time of writing this essay(Department of Health, 2018). However, the public health benefits far outweigh the potential problems associated with My Health Record. The medicine of tomorrow will be based on data and the Australian Government had to introduce a scheme such as My Health Record.

On the topic of data security and privacy, the Australian Government can never completely guarantee patient anonymity; there is always going to be the risk of a data breach. Despite an individual's valuing of their own privacy, however, people arguably care about their health significantly more. As a result, it can be argued that it is up to us as public health professionals and data scientists to do a better job of communicating the benefits of sharing health data, as well as the value it can bring back to the individual.

OLD AND NEW DATA FOR IMPROVING SKIN HEALTH

Professor Steven Tong is an Infectious Diseases Physician with the Victorian Infectious Diseases Service (VIDS) and Co-Head of the Translational and Clinical Research and Indigenous Health cross-cutting disciplines at the Doherty Institute in Victoria. Tong opened his session by acknowledging the Traditional Owners, the Wurundjeri people of the land upon which the symposium took place; he also paid respect to any Indigenous peoples in the audience and acknowledged the use of Indigenous people's data being presented as part of his presentation. This acknowledgement was particularly worthy of mention here; it was clearly appropriate considering the focus of the session was on Indigenous Australian health. While the two previous sessions were primarily focused on big data, Tong spoke to the many potential public health benefits of data linkage using Indigenous health as an example (Olver, 2014).

Data linkage as a method brings information from different sources together; it also collates data about an individual or an entity to create new and often much richer sets of data (Menzies Institute for Medical Research, n.d., para. 1). Data linkage is achieved by assigning a unique identifier to each person across multiple, smaller datasets, which are often called 'orphan datasets.' Within these orphan datasets are a series of links that create connections across all of that individual's personal records. Australia is one of the few countries that has invested heavily in the creation of data linkage facilities and projects. Today, data linkage is predominantly being used for studies of health service outcomes, as well as in epidemiology and needs-based analysis. Data linkage is enabling large-scale studies of whole populations across the healthcare system. As the data is pooled together and the database grows larger, increased statistical power can enable users to explore population-level trends that would have otherwise been unable to be tested. Similar to the debate around NDIS and My Health Record data, the people involved in these types of projects need to understand the benefits from linked health data.

To illustrate the power of data linkage, Tong presented a study he led that pulled together a number of orphan datasets from different communities in the Northern Territory. The diseases of interest were skin infections amongst Indigenous Australian populations, particularly in children.

According to the Menzies School of Health Research (2019, para. 1) in Darwin, childhood skin infections can be extremely serious, with a series of long-lasting and devastating effects recorded. The most serious bacterial infections are *Staphylococcus* and Group A *Streptococcus*, which can lead to Acute Rheumatic Fever and Rheumatic Heart Disease (RHD Australia, n.d., para. 1). The final compiled data set of Tong's study included data gathered at the individual, household and community-level. These findings provided a detailed picture of what was really happening in relation to skin disease prevalence and transmission across the Northern Territory, which would have remained otherwise concealed without data linkage methods.

The presentations demonstrated that there are significant benefits to using big data in healthcare, that is by making it more complete, sharing data and linking data. However, concerns remain about who benefits from these advances in data use particularly around secondary use. Promising models for data sharing such as Dynamic consent (DC) facilitate participant consent and engagement in research over time through an interactive platform (Prictor et. al. 2019). Further work is needed on data governance models, access policies, interoperability, quality assurance and ownership. In addition, micro interventions like technical and consent-related initiatives are limited, in the face of deep seated structural challenges to using and sharing data. Consequently sustained dialogues about how to create ecosystems that incorporate legal protections, collaboration, funding, skill development and new ways of thinking are required (Deetjen et al. 2015).

CONCLUSION

Big data in healthcare is evolving into a promising field; it is proving to be the case that such data collation practices can assist in finding relationships among variables that might otherwise be unrecognisable. It is interesting to think what it might be possible to achieve in healthcare if we could fully harness the power of big data; the Australian population would no doubt benefit from the use of more comprehensive health data. In saying this, we must acknowledge that there are risks attached to such public health gains, notably the concerns around patient privacy and security. Because of these risk factors, it is now the responsibility of leading public health bodies to lobby for a legal framework that protects patient privacy and rights, as well as to effectively communicate the range of benefits of linked health data and the value that it can deliver to the individual

there are significant benefits to using big data in healthcare, that is by making it more complete, sharing data and linking data.

REFERENCES:

Australian Digital Health Agency. (n.d.). My Health Record. Retrieved from https://www.myhealthrecord.gov.au/

Australian Human Rights Commission. (2015). Face the Facts: Disability Rights. Retrieved from https://www.humanrights.gov.au/face-facts-disability-rights

Bogle, A. (2019). My health record: Your questions answered on cybersecurity, police and privacy. Retrieved from https://www.abc.net.au/news/science/2018-07-15/my-health-record-questions-answers-security-privacy-police/9959622

Deetjen, U., Meyer, E. T. and Schroeder, R. (2015), Big data for advancing dementia research: An evaluation of data sharing practices in research on age-related neurodegenerative diseases", OECD Digital Economy Papers, No. 246, OECD Publishing. http://dx.doi.org/10.1787/5js4sbddf7jk-en

Department of Health. (2018). Implementing the Framework to guide the secondary use of My Health Record system data. Retrieved from http://www.health.gov.au/internet/main/publishing.nsf/Content/eHealth-framework

Margo, J. (2018). My Health Record: the benefits and risks explained. Retrieved from https://www.afr.com/lifestyle/health/mens-health/my-health-record-the-benefits-and-risks-explained-20180716-h12rk5

Melbourne Disability Institute. (n.d.). Democratising Disability Data. Retrieved from https://disability.unimelb.edu.au/research/democratising-disability-data

Menzies Institute for Medical Research. (n.d.). What is data linkage? Retrieved from https://www.menzies.utas.edu.au/research/research-centres/data-linkage-unit/what-is-data-linkage

Menzies School of Health Research. (2019). Skin. Retrieved from https://www.menzies.edu.au/page/Research/Indigenous_Health/Skin/

National Disability Insurance Agency (NDIA). (2018). Annual Report. Retrieved from https://www.ndis.gov.au/about-us/publications/annual-report

NDIS. (n.d.). Do you provide supports and services? Find out more Retrieved from https://www.ndis.gov.au/

Olver, I. N. (2014). Linkage data to improve health outcomes. *Medical Journal of Australia*. 200(7), 368-369. doi:10.5694/mja14.00374

Prictor, M., Lewis, M. A., Newson, A. J., Haas, M., Baba, S., Kim, H.,Teare, H. J. A. (2019). Dynamic Consent: An Evaluation and Reporting Framework. *Journal of Empirical Research on Human Research Ethics: JERHRE*, 1556264619887073. https://doi.org/10.1177/1556264619887073

PWC. (2011). Disability Expectations: Investing in a better life, a stronger Australia. Retrieved from https://www.pwc.com.au/industry/government/assets/disability-in-australia.pdf

RHD Australia. (n.d.). What is Acute Rheumatic Fever? Retrieved from https://www.rhdaustralia.org.au/what-acute-rheumatic-fever

Stieb, D. M., Boot, C. R., & Turner, M. C. (2017). *Promise and pitfalls in the application of big data to occupational and environmental health]. BMC public health*, 17(1), 372. https://doi.org/10.1186/s12889-017-4286-8

The Age. (2019). There are real benefits to be had from the My Health Record. Retrieved from https://www.theage.com.au/national/there-are-real-benefits-to-be-had-from-the-my-health-record-20190128-p50u4q.html

United Nations. (n.d. 2006). Convention on the Rights of Persons with Disabilities (CRPD). Retrieved from https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

Zhou, N. (2018). My health record: privacy, cybersecurity and the hacking risk. Retrieved from https://www.theguardian.com/australia-news/2018/jul/16/my-health-record-privacy-cybersecurity-and-the-hacking-risk

ABOUT THE SPEAKERS AND WRITERS

Professor Michele Acuto is an expert on urban politics and international urban planning. Michele is also interim Director of AURIN (the Australian Urban Research Infrastructure Network), a non-resident Senior Fellow of the Chicago Council on Global Affairs and a Senior Fellow of the Bosch Foundation Global Governance Futures Program. Before joining the Faculty, Michele was Director of the City Leadership Lab and Professor of Diplomacy and Urban Theory at University College London, having previously worked as Stephen Barter Fellow of the Oxford Programme for the Future of Cities at the University of Oxford. He also taught at the University of Canberra, University of Southern California, Australian National University and National University of Singapore.

Dr. Gideon Aschwanden is Lecturer in Urban Analytics with a focus on data, learning algorithms and neural networks to evaluate the urban fabric with a focus on health, transportation and economic opportunities. He teaches graduate subjects in Urban Design and Property and leads research projects in the realm of open data, policy and transportation. Before the University of Melbourne, he taught and researched at Princeton University on digital fabrication methods and building systems as part of the CHAOS lab (cooling and heating architecturally optimized systems) in the Andlinger Centre for Energy and the Environment. He has an MSc in Architecture and a doctoral degree in Science from the ETH Zurich. To deepen his knowledge, he worked as a researcher at the Future Cities Laboratory in Singapore and has professional experience in Switzerland, Singapore and the United States.

Amba-Rose Atkinson is a proud Gumbaynggirr, Wakka Wakka, Kabi Kabi and Jinibara woman from the mid north coast of NSW. Amba-Rose began a Master of Public Health degree at the University of Melbourne in 2017, specialising in Global Health with a minor specialisation in Indigenous and Ageing Health. Amba-Rose's passion for the human rights and health rights of her people is further demonstrated by Amba-Rose's active cultural and academic involvement in her community. Throughout her master's degree, Amba-Rose has gone on to assume multiple tutoring roles at the University of Melbourne. Amba-Rose strongly believes in nurturing a community that continues to nurture her. Throughout 2018, Amba-Rose was also a research assistant for Professor Marcia Langton's book, Welcome To Country. All of these experiences have provided continual inspiration and guidance for Amba-Rose's public health career and academic journey. From her studies and life experience, Amba-Rose would like to continue exploring the intersectionality between the environment and Indigenous ways of caring for country, kindling ambition for further study and research, with a strong global, environmental and Indigenous health focus.

Howard Bondell is Professor of Statistical Data Science and Deputy Head of the School of Mathematics and Statistics at the University of Melbourne, and also coordinates the Master in Data Science degree program. He completed his PhD in Statistics in 2005 at Rutgers University in the US, and immediately commenced in a tenure-track position as Assistant Professor at North Carolina State University, then Associate Professor (with tenure) in 2011 and Professor in 2016. In 2017, Prof Bondell was elected Fellow of the American Statistical Association (ASA), the preeminent organisation of both academic and professional statisticians worldwide. He took up the role at the University of Melbourne in 2018. Prof Bondell's research publications are diverse and cover many of the most fundamental areas of statistics, including high-dimensional inference, robustness, model selection, and computational approaches for complex data. He has supervised over 20 PhD students to completion, where they are currently employed in positions in academia, pharmaceutical research, government agencies, as well as high profile tech companies such as Google, Amazon, Facebook, and Apple. Prof Bondell sits on editorial boards for the highest tier journals in statistics, and has been a part of numerous committees in the profession. He is currently Associate Editor for both the Journal of the American Statistical Association and the Journal of the Royal Statistical Society, the two most highly regarded journals in the field of Statistics.

Ishita Chatterjee is an architect who has worked on projects in India and China before shifting to academia. Currently, she is pursuing her PhD along with teaching and working as a research assistant at the University of Melbourne. Her research explores the morphological characteristics of informal settlements by studying its formation process and identifying the socio-political and topographical factors that influence it. When she is not putting informal settlements on the map using satellite imagery she can be found strolling through the alleyways and lanes of different cities around the world.

Darren Clinch: I am a Badimia man from Yamatji country (mid-west of Western Australia). I have been living and working in Melbourne for almost ten years and I currently reside in the western suburbs. I completed a Master of Public Health through Deakin University's Institute of Koorie Education, while working fulltime at the Department of Health and Human Services. I am currently working in the Business Intelligence and Geospatial Support team within the System Intelligence and Analytics branch, after having been in the Aboriginal Health and Wellbeing branch since 2010. My passion is not data for data's sake, but the value that data can bring to a story. My current role involves developing business intelligence solutions for program areas of DHHS wanting to leverage the power of the department vast quantities of data and information for both reporting and monitoring, analytics, and narrative style dashboards. Included in this work is the development of an Aboriginal Information System that can enables users to interpret patterns, trends and relationships within datasets that are presented using 'associative data modelling' visualisations. A key aim for me personally is to promote Indigenous Data Sovereignty and Governance principles and practices within DHHS as per the communique release after the Indigenous Data Sovereignty Summit held in Canberra earlier this year.

Professor Simon Dennis is currently the Director of the Complex Human Data Hub in the Melbourne School of Psychological Sciences and has previously held positions as the Head of School of Psychology at the University of Newcastle, Director of the Cognitive Science Centre at Ohio State University, Associate Professor at the University of Adelaide, Research Professor at the University of Colorado, Boulder, and was a lecturer and senior lecturer in the ARC Centre for Human Factors and Applied Cognitive Psychology, at the University of Queensland. Professor Dennis is a computer scientist by training and has extensive experience in the computational modelling of episodic and semantic memory (Dennis & Humphreys, 2001; Osth & Dennis, 2015; Landuaer, McNamara, Dennis & Kintsch, 2007; Dennis, 2004, 2005). Professor Dennis has been developing and applying passive and active experience sampling methods to understand human experience (Sreekumar, Dennis, Doxas, Zhuang, & Belkin, 2014) and memory (Nielson, Smith, Sreekumar, Dennis & Sederberg, 2015) and has created an extensive data collection, retrieval, visualization and analysis ecosystem provided by Unforgettable Research Services Pty Ltd of which he is the CEO.

Dr Ruth De Souza's interest in data-intensive transformations in health and wellness within the context of informationalization comes from her own research examining how being an active member of one's 'own health care team' through the acquisition of knowledge and skills is a key feature of contemporary health care. Ruth is interested in the advent of the agentic consumer and how this subject position is intensified through digital health, and the health literacy demands of collecting and interpreting data often without the involvement of health professionals. In particular, how these consumer-oriented technologies designed to promote healthy living impact on marginalised communities and how structural inequities in different contexts are inflected in and through technology use.

Professor Anne Kavanagh is the inaugural Chair of Disability and Health at the Melbourne School of Population and Global Health at the University of Melbourne. She is Academic Director of the Melbourne Disability Institute – an interdisciplinary Institute which brings together researchers, government and non-government organisations, and advocates to create systemic change to improve the lives of people with disability. She is Co-Director of the NHMRC Centre of Research Excellence in Disability and Health. Anne is a social epidemiologist and her research focusses on generating high quality policy-relevant evidence for change. She applies high-level quantitative methods to investigate complex social and health problems. Her research spans a range of social determinants including disability, gender, the built environment, socioeconomic position, employment and housing. She is a member of the Victorian Disability Advisory Council and the Independent Advisory Council of the National Disability Insurance Agency. Her research is enriched by her own lived experience as someone with a chronic condition and as a parent of a child with a disability. Anne has a PhD in Epidemiology and Population Health from the Australian National University and is a Fellow of the Australasian Faculty of Public Health Medicine and the Academy of the Social Sciences of Australia.

Professor Marcia Langton AM is an anthropologist and geographer, and since 2000 has held the Foundation Chair of Australian Indigenous Studies at the University of Melbourne. Professor Langton is a Fellow of the Academy of Social Sciences in Australia, a Fellow of Trinity College, Melbourne and an Honorary Fellow of Emmanuel College at the University of Queensland. In 2016 Professor Langton was honoured as a University of Melbourne Redmond Barry Distinguished Professor. In further recognition as one of Australia's most respected Indigenous Academics Professor Langton has in 2017 been appointed as the first Associate Provost at the University of Melbourne.

Julie McLeod is Professor of Curriculum, Equity and Social Change in the Melbourne Graduate School of Education and Pro Vice-Chancellor (Research Capability) at the University of Melbourne. Julie researches in the history and sociology of education, with a focus on youth, gender and inequalities and is currently engaged in several projects that intersect with the digital humanities and the archiving and re-use of qualitative data.

Professor Jodie McVernon is a physician with subspecialty qualifications in public health and vaccinology. She has extensive expertise in clinical vaccine trials, epidemiologic studies and mathematical modelling of infectious diseases, gained at the University of Oxford, Health Protection Agency London and The University of Melbourne. Her work focuses on the application of a range of cross-disciplinary methodological approaches including mathematical and computational models, to synthesise insights from basic biology, epidemiological data and sociological research. These models advance understanding of the observed epidemiology of infectious diseases and inform understanding of optimal interventions for disease control.

Dr Lyndon Ormond-Parker is an ARC Research Fellow in the Indigenous Studies Unit, Centre for Health Equity, Melbourne School Population and Global Health, The University of Melbourne. Dr Ormond-Parker's current ARC-funded research Aboriginal Remote Narrowcast TV and the Audiovisual Archive (2018-2021). Dr Ormond-Parker has a corporate background in communications, public relations, market research and PhD qualifications in anthropology and information technology. His area of expertise relates to Indigenous cultural heritage, information technology and Indigenous communities. He is member of the Australian Heritage Council and member of the Advisory Committee on Indigenous Repatriation. Lyndon was born in Darwin and of Alyawarr decent from the Barkly tablelands region of the Northern Territory.

Dr Tess Ryan is a Biripi woman originating from Taree, New South Wales. Following an early career in government and foster care, Tess moved into academia where she completed a PhD at The University of Canberra focusing on Indigenous women and their experiences of leadership in Australia. Tess was awarded the University Medal in 2013 for her Honours thesis, 'The push/pull indicators of Indigenous political engagement'. Currently Tess works as a Post-Doctoral Fellow with The Poche Centre for Indigenous Health at The University of Melbourne. Her work focuses on Indigenous women, leadership, health and education, and overall representation of Indigenous people in an Australian context.

Dr Soheil Sabri is an Urban Planner and Research Fellow in Urban Analytics at the Centre for Spatial Data Infrastructures and Land Administration (CSDILA) in Melbourne School of Engineering, The University of Melbourne. His research focuses on enabling spatial information and technological innovation in smart urban planning and design to improve urban quality of life.

Dr Tyne Daile Sumner is a Research Fellow in the Digital Studio, Faculty of Arts at the University of Melbourne. Her research operates at the crossroads of poetics, surveillance and big data. She has published essays with Bloomsbury Academic and Australian Literary Studies on topics including surveillance, censorship, poetics and twentieth-century American media. She is currently engaged in new research project, 'Poetry in the Age of Big Data', that critically examines the role of poetic discourse in the context of big data's objective of total algorithmic control.

Dr Vanessa Teague is the chair of the Cybersecurity and Democracy Network at The University of Melbourne. She did her Bachelor's Degree at The University of Melbourne and her Ph.D. in cryptography and game theory at Stanford University. Her research interests are in applied cryptography and online privacy, with interests in electronic voting and other questions relevant to public policy.She was a major contributor to the Victorian Electoral Commission's end-to-end verifiable electronic voting project, the first of its kind to run at a state level anywhere in the world. She recently co-discovered serious security vulnerabilities in the NSW iVote Internet voting system. Her research team has been very active in the analysis of privacy implications of open data and record linkage - most of this is joint work with Dr Chris Culnane and A/Prof Ben Rubinstein. Relevant work includes a demonstration that an open MBS-PBS dataset allowed re-identification of doctors and patients. They have also worked with ABS on the analysis and design of privacy-preserving linkage processes and with Transport NSW and the Office of the Victorian Information Commissioner on the analysis and design of data privacy techniques.

Associate Professor Steven Tong is an infectious diseases physician with the Victorian Infectious Diseases Service and Co-Head of the Translational and Clinical Research and Indigenous Health cross-cutting disciplines at the Doherty Institute. He spent 10 years in Darwin before moving Melbourne to join the Doherty Institute in 2016. His research interests include skin pathogens (Staphylococcus aureus, Group A Streptococcus), hospital infections, Indigenous health, viral hepatitis and influenza. His passion is to apply cutting edge science to address clinically driven questions.

Kyle Turner is a Lecturer in Indigenous Health within the School of Population and Global Health at the University of Melbourne. Along with teaching, he has research commitments in the area of dementia prevention in older Aboriginal people at high risk. Prior to this, Kyle completed his doctorate in public health at the University of Oxford and the Master of Applied Epidemiology at the Australian National University.

Dr Fiona Tweedie first became interested in the intersection between data, technology and privacy while working at the Office of the Australian Information Commissioner on issues ranging from Google Streetview to open data. She subsequently organised the Melbourne node of GovHack, the world's largest civic hackathon, in 2013 and 2014, and served on the board of Open Knowledge Australia from 2013 to 2016. Her current role at the University of Melbourne includes building ethical frameworks for data analysis that seek to centre vulnerable groups and people in project design. She is interested in digital humanities and supporting cross-disciplinary collaborations. She holds a PhD in Roman History.

Karin Verspoor is a Professor in the School of Computing and Information Systems and Deputy Director of the Health and Biomedical Informatics Centre at the University of Melbourne. Trained as a computational linguist, Karin's research primarily focuses on extracting information from clinical texts and the biomedical literature using machine learning methods to enable biological discovery and clinical decision support. Karin held previous posts as the Scientific Director of Health and Life Sciences at NICTA Victoria Research Laboratory, at the University of Colorado School of Medicine, and Los Alamos National Laboratory. She also spent 5 years in start-ups during the US Tech bubble, where she helped design an early artificial intelligence system.

ISBN 978-0-7340-5596-5